

AD-A041 687

ILLINOIS UNIV AT URBANA-CHAMPAIGN COORDINATED SCIENCE LAB F/G 9/5
DESIGN CONSIDERATIONS AND TRADE-OFFS IN MOS/LSI.(U)

APR 77 A D GANT

DAAB07-72-C-0259

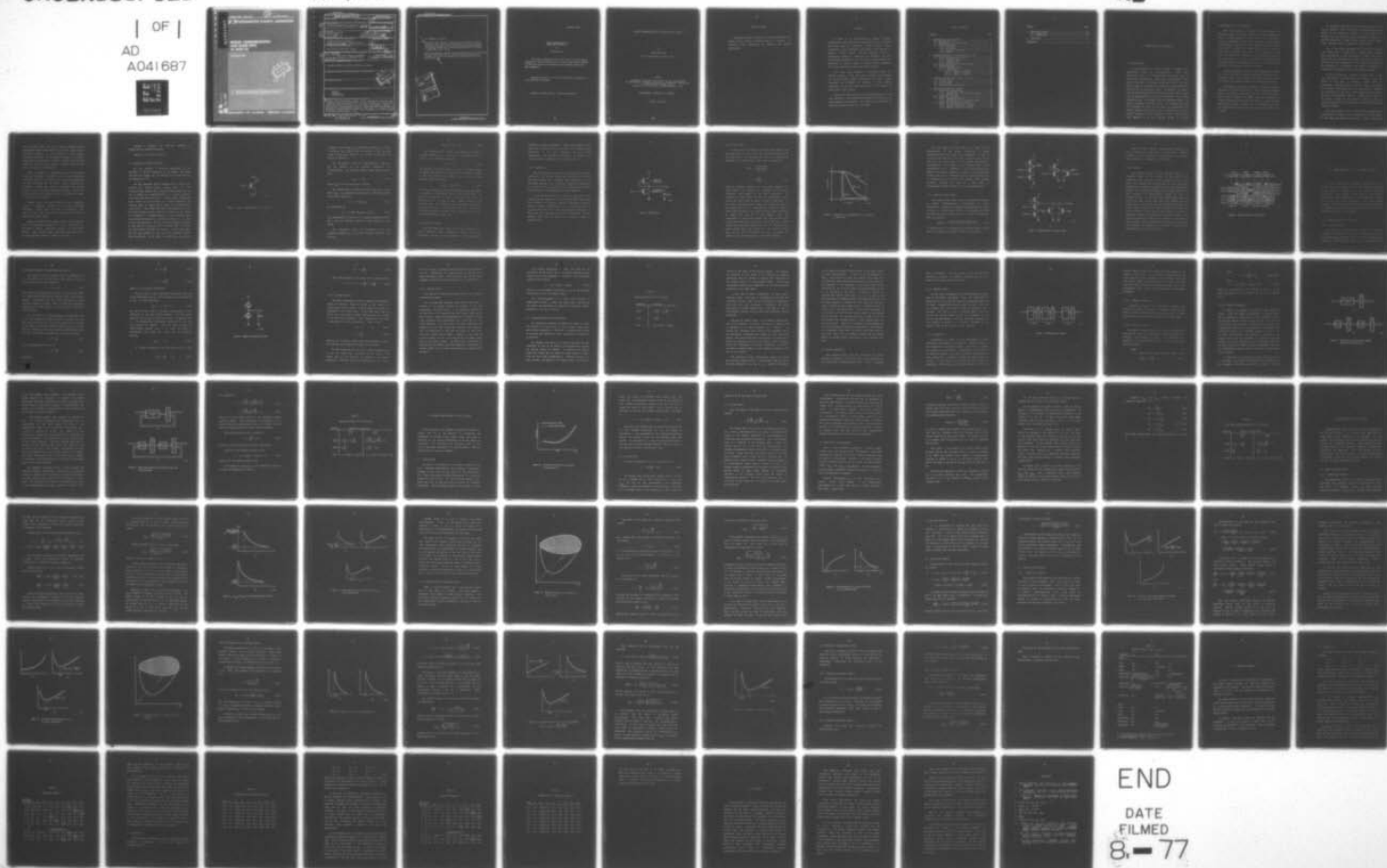
UNCLASSIFIED

R-764

NL

| OF |

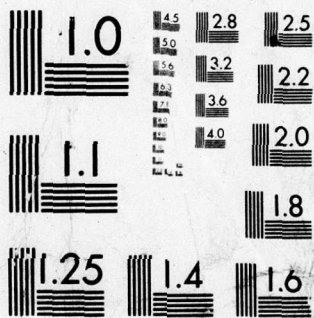
AD
A041 687



END

DATE
FILMED

8-77



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A041687

CSL COORDINATED SCIENCE LABORATORY

DESIGN CONSIDERATIONS AND TRADE-OFFS IN MOS/LSI

ALAN DALE GANT



APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED.

AD No. _____
DDC FILE COPY

UNIVERSITY OF ILLINOIS - URBANA, ILLINOIS

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle)		5. TYPE OF REPORT & PERIOD COVERED
(6) DESIGN CONSIDERATIONS AND TRADE-OFFS IN MOS/LSI.		Technical Report
7. AUTHOR(s)		6. PERFORMING ORG. REPORT NUMBER
(10) Alan Dale Gant		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS		DAAB-67-72-C-0259
Coordinated Science Laboratory University of Illinois at Urbana-Champaign Urbana, Illinois 61801		MCS-73-03488 AOL
11. CONTROLLING OFFICE NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Joint Services Electronics Program		(11)
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		12. REPORT DATE
(14) R-7643 KILU-ENG-77-2211		Apr 1977
		13. NUMBER OF PAGES
		84
		15. SECURITY CLASS. (of this report)
		UNCLASSIFIED 92p.
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)		
Approved for public release; distribution unlimited		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
MOS/LSI Pipelining Optimal Design		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)		
<p>In order to be cost-effective, digital circuits implemented in MOS large scale integration (LSI) are designed with close regard to optimality. Three criteria - power dissipation, chip area (cost), and speed - are balanced to produce a final circuit. Current practices involve a significant amount of simulation to determine the most satisfactory trade-offs. This project is concerned with the prediction of optimal values of certain design parameters for a given function of the above three criteria.</p>		

DD FORM 1473
1 JAN 73

EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

097700

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

20. ABSTRACT (continued)

This paper first develops a simplified model for MOS/LSI circuits. Using this model, equations are derived for circuit speed, power dissipation, and area. These equations are expressed as functions of the channel dimensions of the MOS transistors and of the degree of pipelining of the circuit.

These three equations, and their pairwise products, are then optimized. The relationships between the equations and the parameters are examined. The effects and trade-offs of particular design choices are described.

Approved for	White Section <input checked="" type="checkbox"/>	Buff Section <input type="checkbox"/>
RTIS		
UNANNOUNCED		
JUSTIFICATION		
BY	DISTRIBUTION/AVAILABILITY CODES	
Dist.	AVAIL. and/or SPECIAL	
A		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

UILU-ENG 77-2211

DESIGN CONSIDERATIONS AND
TRADE-OFFS IN MOS/LSI

by

Alan Dale Gant

This work was supported in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy and U.S. Air Force) under Contract DAAB-07-72-C-0259 and in part by the National Science Foundation under Grant MCS 73-03488 A01.

Reproduction in whole or in part is permitted for any purpose of the United States Government.

Approved for public release. Distribution unlimited.

DESIGN CONSIDERATIONS AND TRADE-OFFS IN MOS/LSI

BY

ALAN DALE GANT

B.S., University of Texas, 1974

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1977

Thesis Adviser: Professor E. S. Davidson

Urbana, Illinois

ACKNOWLEDGEMENT

The author wishes to express his sincere gratitude and thanks to his advisor, Professor Edward S. Davidson. Dr. Davidson's many suggestions and patience are greatly appreciated.

ABSTRACT

In order to be cost-effective, digital circuits implemented in MOS large scale integration(LSI) are designed with close regard to optimality. Three criteria - power dissipation, chip area(cost), and speed - are balanced to produce a final circuit. Current practices involve a significant amount of simulation to determine the most satisfactory trade-offs. This project is concerned with the prediction of optimal values of certain design parameters for a given function of the above three criteria.

This paper first develops a simplified model for MOS/LSI circuits. Using this model, equations are derived for circuit speed, power dissipation, and area. These equations are expressed as functions of the channel dimensions of the MOS transistors and of the degree of pipelining of the circuit.

These three equations, and their pairwise products, are then optimized. The relationships between the equations and the parameters are examined. The effects and trade-offs of particular design choices are described.

TABLE OF CONTENTS

Chapter	Page
1. INTRODUCTION AND BACKGROUND	1
1.1 Introduction	1
1.2 Operation of MOS Transistors	5
1.3 Digital MOS Gates	8
1.3.1 Inverters	9
1.3.2 Beta Ratio	11
1.3.3 Multiple Input Gates	13
1.3.4 Layout	15
2. CHARACTERIZATION OF A SIMPLIFIED MOS MODEL	17
2.1 Characterization of Inverters	17
2.1.1 Inverter Speed	17
2.1.2 Inverter Power	21
2.1.3 Inverter Area	22
2.2 Characterization of Complex Gates	23
2.3 Storage Registers	26
2.3.1 Register Model	27
2.3.2 Pipelining	27
2.3.2.1 Degree 0 Pipeline	29
2.3.2.2 Degree 1 Pipeline	29
2.3.2.3 Degree m Pipeline	30
3. OPTIMUM CIRCUIT SPEED, POWER, AND AREA	36
3.1 Circuit Area	36
3.2 Circuit Power	36
3.3 Circuit Speed	39
3.4 Beta Ratio Considerations	40
4. OPTIMA INVOLVING TWO CRITERIA	45
4.1 Speed and Power Optima	45
4.1.1 Speed-Power Product	45
4.1.2 Minimum Power for Specified Speed	50
4.2 Area-Power Product	55
4.3 Speed and Area Optima	56
4.3.1 Speed-Area Product	56
4.3.2 Minimum Area for Specified Speed	62
4.4 Optima for Combinational Logic	68
4.4.1 Optimum Speed-Power Product	68
4.4.2 Optimum Power-Area Product	68
4.4.3 Optimum Speed-Area Product	69

Chapter	Page
5. Numerical Examples	72
5.1 Example One	73
5.2 Example Two	75
6. CONCLUSION	81
REFERENCES	84

1. INTRODUCTION AND BACKGROUND

1.1 Introduction

In the design of a digital circuit or system, many alternatives exist for implementation. Large Scale Integration (LSI) is an implementation technology which is being used increasingly. Metal Oxide Semiconductor (MOS) devices have been a dominant technology for LSI for many years, boasting extremely high densities (gates/integrated circuit) and low production costs. In addition, in density and cost-effectiveness MOS technology has maintained a rapid rate of improvement ever since its inception. The process considered here is N-channel MOS with depletion loads. All logic is assumed to be ratio logic. The results also hold for P-channel but they do not apply to complementary MOS (CMOS), because ratio logic, while predominant in N and P channel circuits, is not utilized in CMOS. The depletion load appears to be the dominant choice in current

single-channel (P or N) devices.

Once the decision to implement a given circuit in MOS/LSI is made, the circuit designer has many decisions to make. For some systems, the logic will not fit economically onto one integrated circuit (IC) and must be partitioned appropriately. When the circuit designer is finally ready to implement a given logic circuit within an IC, there are still many parameters with which to be concerned. These fall into three major categories - 1) parameters related to the processing to be used, 2) parameters related to the intended operating environment of the IC, and 3) parameters over which the designer has direct control.

Process parameters dictate to the designer the minima and maxima imposed to achieve respectable yields on the IC production line. These include such things as minimum metal line widths and minimum transistor-to-transistor spacing.

Environmental parameters dictate the required interface from the IC to the outside world. They include temperature ranges, and signal requirements for both input and output. Power supply voltages are often specified, although the circuit designer may have some discretion in order to meet performance criteria. Process and environmental parameters are constraints and bounds from the circuit designer's point of view.

The remaining parameters are the primary tools of the circuit designer. Thus, the number of parameters which are totally at the designer's disposal is quite small. In MOS/LSI, they are effectively only the channel dimensions of the MOS transistors. These directly affect circuit speed, power, and area, which is related to circuit cost.

All of the above parameters are utilized to achieve some design goal - minimum cost, maximum speed, minimum power dissipation, or some combination of these. This paper deals with the relationships between the various goals as evidenced by their dependence on the channel dimensions of the transistors. In addition, the effect of pipelining is examined as an additional tool of the circuit designer.

As the semiconductor industry considers much of its data proprietary, raw numbers are not available. It is the intent of this paper to provide a first order, or approximate, characterization of MOS/LSI. Thus, the results will include coefficients which are unevaluated. Comments and conclusions on the form of these results are presented, including the trade-offs implied. Readers with access to the actual figures in the industry should be able to calculate values for the coefficients and obtain first order numerical results.

The logic design of the circuit to be implemented is considered to be frozen, as the optimality of a logic design is not within the scope of this paper. Given a particular

logic circuit, then, how can a circuit designer choose appropriate channel dimensions and, possibly, the degree of pipelining desired? In the remainder of this chapter, necessary background in MOS digital devices is provided. These topics include MOS transistor operation, MOS digital inverters, the Beta ratio concept, multiple input digital gates, and layout considerations.

Then, in Chapter 2, a simplified model of MOS digital devices is developed. The model is first used to describe a single MOS inverter. The speed, area, and power dissipation are then specified for the model. Next, arguments are presented to extend the model to represent more complex circuits, up to an entire integrated circuit. Models for storage devices are developed, which also model the degree of pipelining of a circuit.

Chapter 3 uses the model developed in the preceding chapter to ascertain optima for the three operating parameters - speed, power, and area, individually. With these in mind, the Beta ratio constraints are used to simplify the model further.

Chapter 4 presents more complex optima, those involving two of the operating parameters at a time. Among these are speed-power product, speed-area product, and power-area product. Also, optimum power dissipation and optimum area are each examined assuming a bound on the speed.

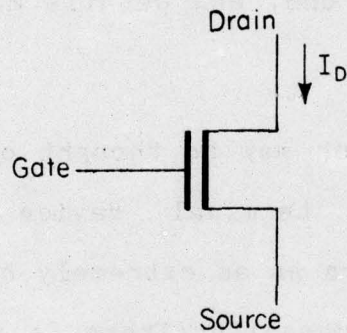
Chapter 5 presents two numerical examples to demonstrate the methods developed.

Chapter 6 is a short conclusion.

1.2 Operation Of MOS Transistors

For our purposes, a detailed description of the operation of the MOS transistor is not needed. The model used here is a simple one, and details may be obtained from the references.^{1,2,3}

The MOS transistor may be thought of as a voltage in, current out, three terminal device (Fig. 1). The transistor gate appears as an extremely high resistance, and may be modeled as a capacitor (There is a certain ambiguity associated with the word "gate". In this paper, "transistor gate" refers to the terminal of a MOS transistor, and "digital gate" or "logic gate" refer to a combinational logic gate.). The source and drain are identical, but the one whose voltage is most negative is labeled as the source (in N-channel). A voltage applied to the transistor gate must exceed the source by a threshold voltage, V_t , or more to turn the device on. If V_{gs} , the gate-to-source voltage, is less than V_t , the device is off and the drain and source are electrically isolated (no current). On the other hand, the more V_{gs} exceeds V_t , the lower the impedance between drain and source. It is useful to define V_{ton} , the turn-on



FP-5418

Figure 1. Symbolic Representation of MOS Transistor

voltage, as the amount V_{gs} exceeds the threshold, V_t . Thus, a negative V_{ton} implies that the device is off, and the value of V_{ton} , when positive, is a measure of how hard the device is turned on.

The transistor's operating characteristics, when on, may be modeled in two regions, saturation and non-saturation. The boundary between these regions occurs when

$$V_{ton} = V_{ds} , \quad (1-1)$$

where V_{ds} is the drain-to-source voltage.

The region known as saturation occurs when $V_{ton} < V_{ds}$, while $V_{ton} > V_{ds}$ implies non-saturation. The drain current (=the source current) is

$$I_d = K' \frac{W}{L} V_{ton}^2 \quad (1-2)$$

for saturation and

$$I_d = K' \frac{W}{L} (2V_{ton}V_{ds} - V_{ds}^2) \quad (1-3)$$

for non-saturation.⁴ Here K' is a constant coefficient and W and L are the channel dimensions (width,length) of the device.

For transistors which are enhancement mode, the threshold voltage, V_t , is a positive non-zero quantity. So, for V_{ton} :

$$V_{ton_e} = V_{gs} - V_t \quad (1-4)$$

For depletion mode devices, the threshold voltage is actually negative, so a V_{gs} of zero is already in the on region. It is customary to define

$$V_p = -V_t \quad (1-5)$$

for depletion mode transistors. V_p , the pinch-off voltage, represents the amount of voltage which is turning the device on when the gate-to-source voltage, V_{gs} , is zero. So for depletion mode transistors:

$$V_{ton_d} = V_{gs} + V_p \quad (1-6)$$

A zero V_{gs} results in a positive non-zero V_{ton} so the device is normally on (enhancement devices are normally off). It would be necessary to apply a negative V_{gs} equal in magnitude to V_p to turn it off. However, this is not normally done. The depletion mode transistor as used in this paper has $V_{gs} = 0$, so it never turns off. It may then be used as a resistive device, whose resistance varies with its drain-to-source voltage, V_{ds} .

1.3 Digital MOS Gates

Each MOS digital gate, using ratio logic, consists of a single depletion mode transistor, called the load transistor, and one or more enhancement mode transistors,

called the driver transistors. Ratio logic refers to the fact that a logical zero is the result of a voltage divider consisting of the load transistor and the driver transistors. The simplest configuration, an inverter, has just one driver transistor and is shown in Fig. 2.

1.3.1 Inverters -

The load transistor, since it is depletion mode, always remains on. The driver, though, may be turned on or off by its gate voltage, V_{in} . Actually, the transistors are all continuous devices. For a given V_{in} , V_{out} may be determined by setting the two I_d currents equal (taking into account the saturation or non-saturation of each device).⁵

A more intuitive approach is taken in this paper. The drain-to-source impedance of the depletion-mode load device can be viewed as a somewhat constant resistance, certainly within an order of magnitude or so. On the other hand, as V_{in} goes from zero to $+V$, the drain-to-source impedance of the driver will decrease many orders of magnitude. The two devices, then, may be viewed as a voltage divider to determine V_{out} .

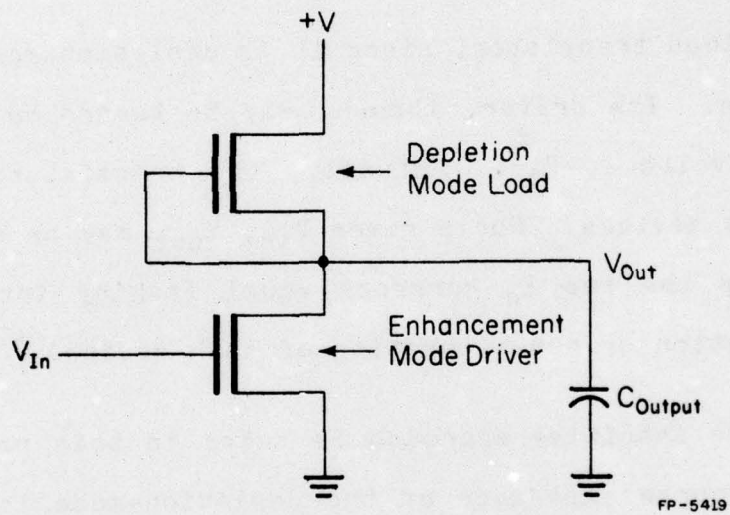


Figure 2. MOS Inverter

1.3.2 Beta Ratio -

A useful tool for denoting the relationship between the two resistances of ratio logic (load and driver sections) is the Beta ratio. It describes the static or steady-state operation of a digital gate. Beta ratio is defined as

$$\text{Beta} = \frac{\left(\frac{W}{L}\right)_{\text{driver}}}{\left(\frac{W}{L}\right)_{\text{load}}} \quad (1-7)$$

$$= \frac{W_d L_l}{W_l L_d} \quad (1-8)$$

Since W/L appears explicitly in the current equation for each device, Beta measures the "strength" of the driver transistor with respect to the load transistor. For instance, for small values of Beta, the "resistance" of the driver may never become as small as that of the load. In this case, the output would remain near $+V$ for any value of V_{in} between zero and $+V$. Alternatively, if Beta is large, a very slight increase in V_{in} , just above V_t , would cause V_{out} to change from near $+V$ to near ground. Here, the "resistance" of the driver becomes very much less than that of the load with only a very small value for V_{ton} . These results are graphically depicted in Fig. 3. Note that V_{out} never changes until V_{in} exceeds V_t . This occurs because V_{ton} for the driver is negative for $V_{in} < V_t$, making the driver off and its "resistance" practically infinite.

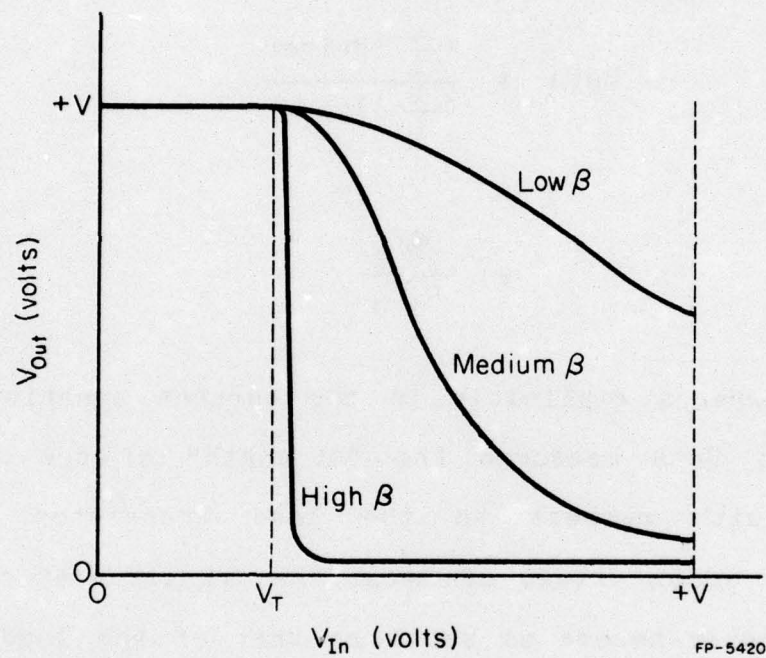


Figure 3. Inverter Transfer Characteristic As A Function of Beta-ratio

The net impact of Beta ratio is a bound on the relationships of the channel dimensions of driver transistors relative to their load transistors. If the W/L for the driver section is too small relative to the W/L for the load, the logic gate will be unable to achieve a satisfactorily low voltage or zero output (even in the steady-state or D.C. condition). Since there are physical minima for channel dimensions imposed by the production process, Beta may be increased beyond a certain amount only by increasing W_d or L_1 above their minima. This increase obviously increases the area of a logic gate, so unreasonably high Beta values are to be avoided as well.

1.3.3 Multiple Input Gates -

More complex logic gates than the inverter are achieved by making series-parallel combinations of driver transistors. Examples of possible configurations are shown in Fig. 4. The circuit in Fig. 4a represents the NAND function for positive logic while Fig. 4b is a NOR circuit. Fig. 4c realizes

$$\text{Output} = \overline{((I_{n1} \wedge I_{n2}) \vee I_{n3} \vee I_{n4})} \quad (1-9)$$

In other words, 4c is a three-input NOR logic gate in which one of the inputs is actually the AND of two inputs.

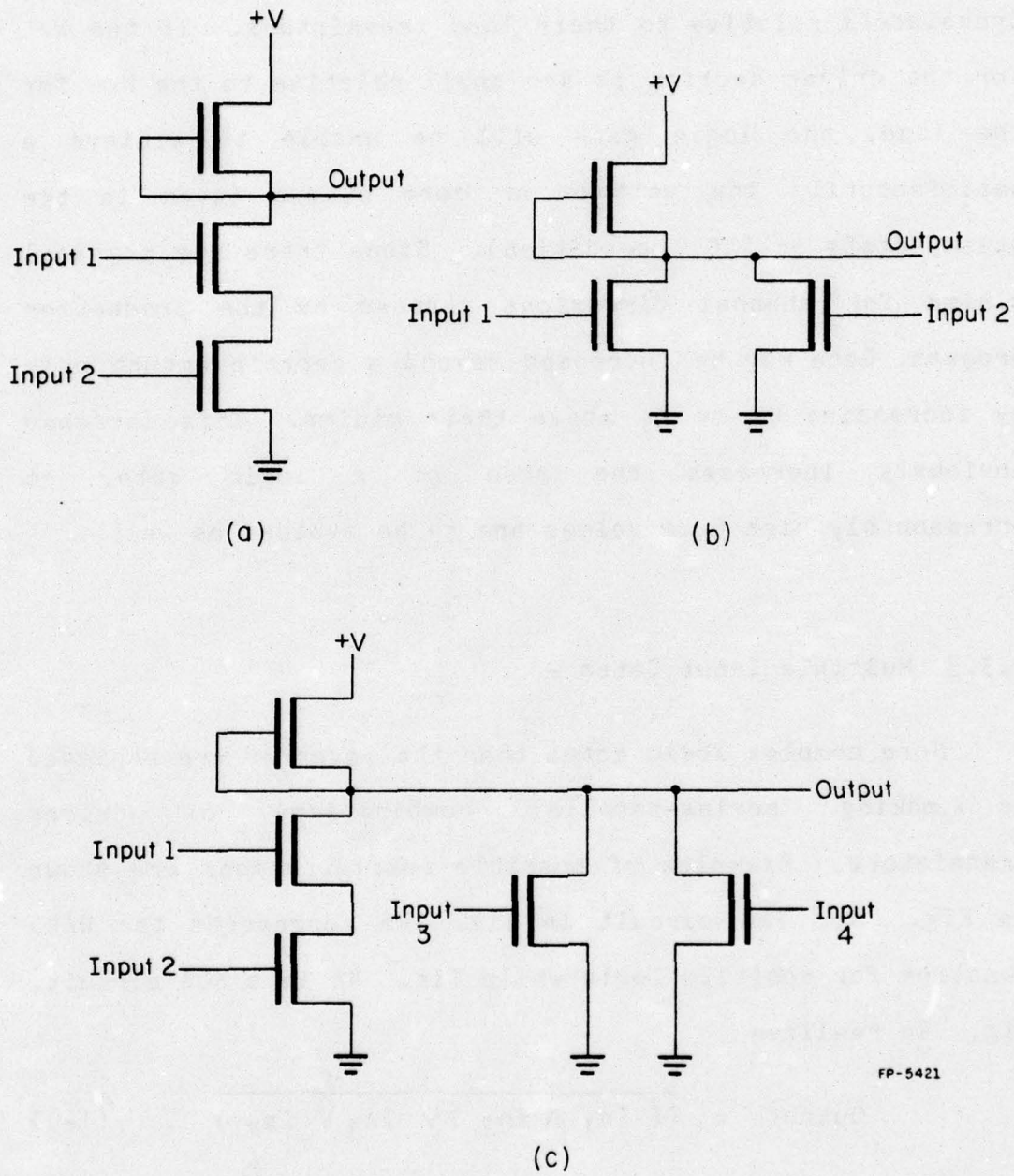


Figure 4. Representative MOS Digital Gates

These and other similar configurations comprise the usual combinational logic gates in ratio logic. Although bridges and other more exotic configurations are possible, their occurrence is assumed to be infrequent.

1.3.4 Layout -

The physical layout of the digital gates is an important topic, as it relates to the area.^{6,7} The transistors are basically surface devices, so the logic gates and circuits may be described in two dimensions. Furthermore, a circuit is usually laid out in a bus-oriented fashion to facilitate distribution of power, ground, and logic signals. A representative section of logic is pictured in Fig. 5. In this figure, four inputs enter the section of logic on the left, while four outputs and two inputs exit on the right, for use in succeeding logic. The dashed regions represent the areas occupied by the load and driver sections of each digital gate, while an X denotes a connection between an interconnect line and a digital gate. As can be seen, much of the area is occupied by the busing of logic signals, power, and ground to appropriate logic gates. The remainder is occupied by the actual transistors and interconnections of each logic gate.

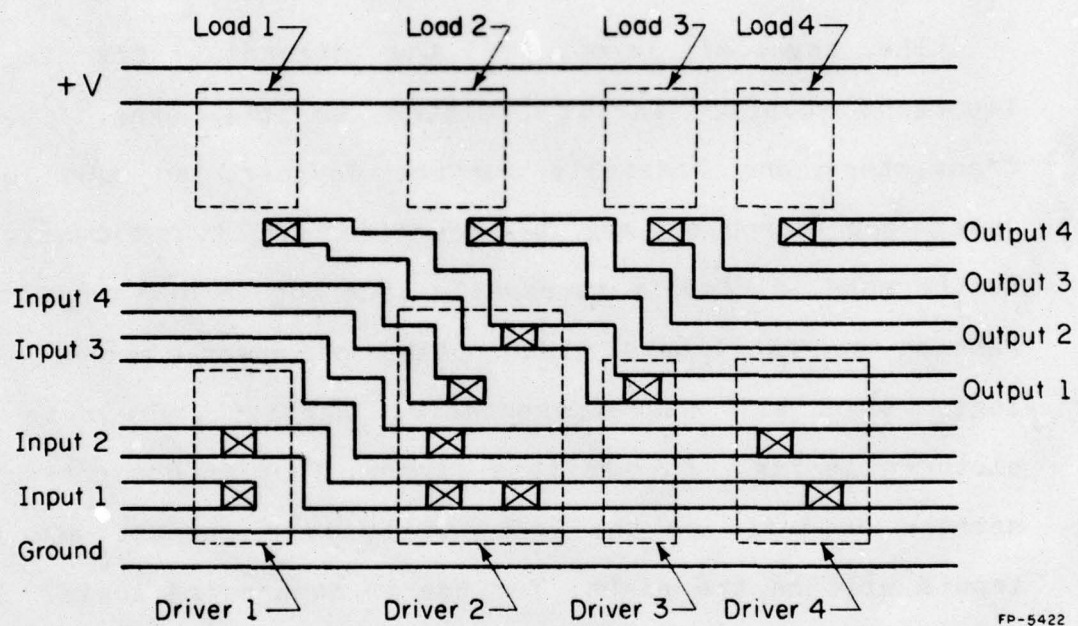


Figure 5. Typical Section of Logic Layout

2. CHARACTERIZATION OF A SIMPLIFIED MOS MODEL

In this chapter the simplified description of MOS devices is used to develop a set of equations which describe three operating functions of the channel dimension parameters: device speed, circuit area, and device power dissipation. These functions are first devised for a single MOS inverter and then extended to model more complex devices. Finally, the characterization of storage elements, or registers, is introduced and the equations are altered to reflect the effects of pipelining.

2.1 Characterization Of Inverters

2.1.1 Inverter Speed -

The speed of semiconductor devices is generally presented by describing the inverse of speed, time, as in the speed-power product which has the units of watt-sec (not watt/sec). Therefore, this paper uses a function related to

device delay time to characterize device speed.

The speed of an MOS inverter could be represented by the rise time or by the fall time of the device. The value chosen here is the sum of these two:

$$T = T_r + T_f . \quad (2-1)$$

This choice is made since the values of rise and fall times are usually quite different, so neither alone is an accurate measure. Furthermore, all MOS logic gates have a single inversion embedded in them. Thus a signal which propagates through two or more levels of logic has both rising and falling delays.

The rise time of a digital inverter is dependent only on the load transistor, since the driver transistor has just been turned off (The time to turn the driver off is included in the time to turn the previous digital gate on.). Given that the transistor is driving a purely capacitive load, the rise time is inversely proportional to the drain current:

$$T_r \propto \frac{1}{I_1} . \quad (2-2)$$

Also, from equations 1-2 and 1-3,

$$I_1 \propto \frac{W_1}{L_1} . \quad (2-3)$$

Therefore,

$$T_r \propto \frac{L_1}{W_1} \quad (2-4)$$

or

$$T_r = k_1 \frac{L_1}{W_1} , \quad (2-5)$$

where k_1 is a constant coefficient.

The fall time is more complicated since both load and driver transistors are on. This situation may be modeled as in Fig. 6. This means that

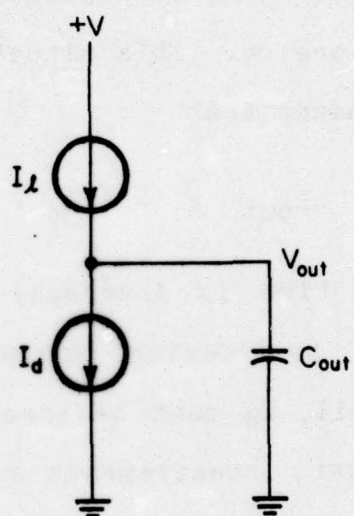
$$I_{out} = - (I_d - I_1) \quad (2-6)$$

and that the fall time is inversely proportional to the difference of the two device currents. For the device output to fall at all, I_d must be greater in magnitude than I_1 . The steady-state requirements mentioned in the Beta ratio section of Chapter 1 are such that I_d must be significantly greater than I_1 in order to obtain a satisfactory zero level out. Thus, for all practical purposes,

$$I_{out} = - I_d . \quad (2-7)$$

In a manner analogous to the rise time, the fall time is

$$T_f \propto \frac{1}{I_d} , \text{ or} \quad (2-8)$$



FP-5423

Figure 6. Model for Inverter Fall Time

$$T_f = k_2 \frac{L_d}{W_d} . \quad (2-9)$$

With these equations, the total time is represented by

$$T = T_r + T_f = k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d} . \quad (2-10)$$

2.1.2 Inverter Power -

The power dissipation of an MOS inverter is relatively simple to evaluate. When the driver is off, there is a virtual open circuit between +V and ground, so no power is dissipated in the static case. On the other hand, when the driver is on, the output voltage is near ground (as a result of the Beta ratio criterion). Thus, almost all of the power is dissipated in the load transistor. Power is then

$$P \propto I_1 , \text{ or} \quad (2-11)$$

$$P = k_3 \frac{W_1}{L_1} , \quad (2-12)$$

where k_3 is a constant coefficient which includes a factor for the duty cycle (percent on time) of the device.

The above equation for the static power dissipation is in fact the total power dissipated by the device. The output load on the inverter is assumed to be purely capacitive. Therefore during any complete 0 to 1 to 0 (or 1

to 0 to 1) cycle, the total energy supplied to the capacitor is zero. Effectively, the capacitor may be removed for a power consumption model. Then the logic gate in the model switches in zero time and the power is as in equation 2-12.

2.1.3 Inverter Area -

The area of an MOS/LSI integrated circuit is a useful cost-related measure.

For a single MOS inverter, only part of the area is related to the channel widths and lengths of the two transistors. As shown in Fig. 5, a significant portion of the area of an IC is occupied by power supply lines and by signal lines. These area components are a function of the logic equations being implemented and of the skill of the layout designer. For this paper, the logic equations are assumed to be in their final form, hopefully near optimum. Also, the assumption is made that the layout is well done, with very little wasted space. Changes in the area of certain portions of the IC are assumed not to affect other areas to any great extent. In other words, the layout is considered to be plastic such that changes or alterations in certain places do not cause inefficiencies or wasted space elsewhere.

With these assumptions in mind, the area can be considered as the sum of the area directly affected by the channels and of the remainder, or overhead. This assumption results in the formula

$$A = k_4 + k_5 W_1 L_1 + k_6 W_d L_d, \quad (2-13)$$

where k_4 is the overhead area, and k_5 and k_6 are constant coefficients of the two channel areas.

The characterization of a single MOS inverter is summarized in Table 1. Here, the speed, power, and area of the device are represented as functions of the four channel dimensions - W_1 , W_d , L_1 , and L_d .

2.2 Characterization Of Complex Gates

The parametric equations presented to model a single MOS inverter are also valid for more complex gates, with some further assumptions. After all, a complex gate is only an inverter with additional driver transistors in series or in parallel.

For example, the speed of a complex gate may be made comparable to that of an inverter by appropriately picking the channel widths and lengths. In general, the channel widths and lengths may be chosen to create faster or slower rise and fall times independently. Whatever decision is made, however, the speed of the complex gates can be easily

Table 1.

Characterization of MOS Inverter

Function	Equation
Speed	$k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d}$
Power	$k_3 \frac{W_1}{L_1}$
Area	$k_4 + k_5 W_1 L_1 + k_6 W_d L_d$

related to the speed of the single inverter. For example, the doubling of the speed of the inverter by changing dimensions could be matched by changing the dimensions of the complex gates by a proportional amount. This linear relationship results from dependence on the same operating equations by all logic gates.

In the case of power dissipation, an even simpler relation exists. The power is dependent only on the load transistor, and all logic gates, regardless of complexity, have only one load transistor. Therefore, the power dissipated in a given complex gate is the same as that of the standard inverter, with the same load transistor channel dimensions.

The area of complex gates is, in general, greater than that of an inverter. In effect, the area of a complex gate is composed of the same components as in Table 1. k_u remains the "overhead" area, which may be increased due to the additional interconnection needed for the additional inputs. The area of the driver transistors is a multiple of the driver transistor area for an inverter, with a possible factor for differences in channel dimensions. As before, the area of complex gates is also linearly related to the area of an inverter.

The linearity of the relationships between all logic gates yields a useful result. A percentage change in any of the four parameters - W_1 , W_d , L_1 , L_d - applied universally

to all gates on a given IC would result in new speed, power, and area specifications. In fact, if the coefficients in Table 1 are altered to model the average logic gate on an IC, then they would describe the operation of the entire IC. For instance, a 10% increase in the W_d of every driver transistor on the IC would alter the specifications of the IC as predicted by the Table 1 equations, modified to describe the average logic gate. W_d , W_l , L_l , and L_d now each describe the average channel dimensions on the IC. In the speed equation, k_1 remains the same, while k_2 includes a factor related to the average number of average drivers in series, i.e., the slowest path. In the power equation, k_3 now includes a factor for the average number of load devices in the ON state. In the area equation, k_4 adds all of the interconnect and periphery area - output bonding pads and off-chip drivers. This area is related to environmental and processing parameters which are unavailable to the designer. k_5 is unchanged since there is exactly one load device on each logic gate, and k_6 includes a factor for the average number of average driver transistors in the average logic gate.

2.3 Storage Registers

This section deals with the insertion of storage elements or registers. Up to this point, the discussion has been restricted to combinational logic. First, a register

model is developed. Then this model is used to extend the parametric equations to sequential machines and to more general multi-stage pipeline machines.

2.3.1 Register Model -

In MOS, many implementations of storage elements exist, such as two-phase, four-phase, ratio, and ratio-less.^{8,9} Rather than try to model the specific differences among all of these, a generalized register model is used. Given any one register type, its delay or speed is assumed to be a constant. Its area is also a constant which takes into account the average number of bits in a register and the area per storage element. Likewise, the power dissipated in each register is modeled as a constant, dependent on the register type and the average bits per register.

2.3.2 Pipelining -

Pipelining is a style of logic architecture in which computation throughput is increased by allowing overlap in the processing of successive tasks. The logic is divided into stages, and the degree of pipelining, m , is equal to the number of stages. Each stage consists of some amount of logic followed by a register to buffer the stage output, as in Fig. 7. The logic in each stage may be of arbitrary complexity. The registers are generally gated or clocked at

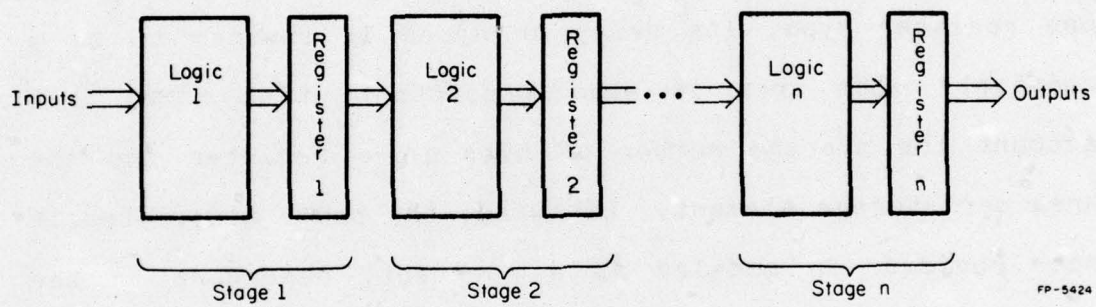


Figure 7. Pipelined Logic Machine

regular intervals which are greater than the maximum of the delays of the stages. In a pipelined implementation, new input operands may be presented at each register clock time and outputs appear at each register clock time. The total time from input to output is generally greater than in the nonpipelined case, but the system throughput may be greatly increased.

2.3.2.1 Degree 0 Pipeline -

A pipeline of degree 0 has, effectively, 0 stages. The pipeline of degree 0 is taken to represent an MOS circuit which is entirely combinational with no registers. For this case, the parametric equations in Table 1 are sufficient.

2.3.2.2 Degree 1 Pipeline -

The degree 1 pipeline consists of a single (arbitrarily complicated) stage of combinational logic followed by a register. The pipeline of degree 1 is analogous to a sequential or finite-state machine. In this case, the equations become:

Speed:

$$T = \text{Combinational delay} + \text{Register delay} \quad (2-14)$$

$$= k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d} + k_7, \quad (2-15)$$

Power:

$$P = k_3 \frac{W_1}{L_1} + k_8, \text{ and (2-16)}$$

Area:

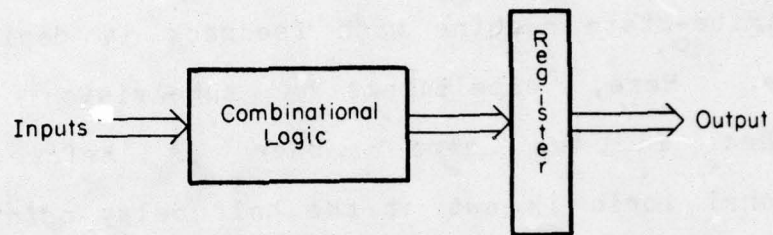
$$A = k_4 + k_5 W_1 L_1 + k_6 W_d L_d + k_9, \quad (2-17)$$

where k_7 is the delay of the register in use, k_8 is the power dissipation in the register, and k_9 is the register area.

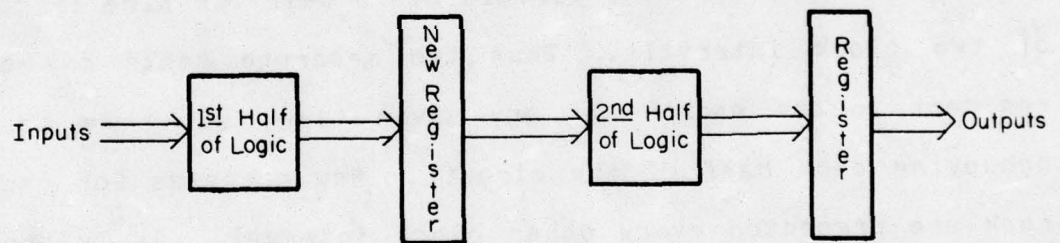
2.3.2.3 Degree M Pipeline -

In general, a pipeline may consist of any number of stages. Also, the degree of pipelining of a given circuit may be increased by subdivision in many cases.^{10,11} For a finite-state machine (degree 1) such as Fig. 8a, where the combinational logic consists of several levels of logic, the combinational section may be separated at approximately the mid-point of delay, and a register inserted as in Fig. 8b. The total delay through the machine is increased by one register delay, but the throughput is definitely increased because new input operands may be introduced after the previous inputs have passed through approximately one-half the combinational delay instead of the entire combinational delay as before.

In actuality, the subdivision process need not be done in two parts. Any number of registers may be inserted, up to a maximum of one after each level of logic. In these



(a)



(b)

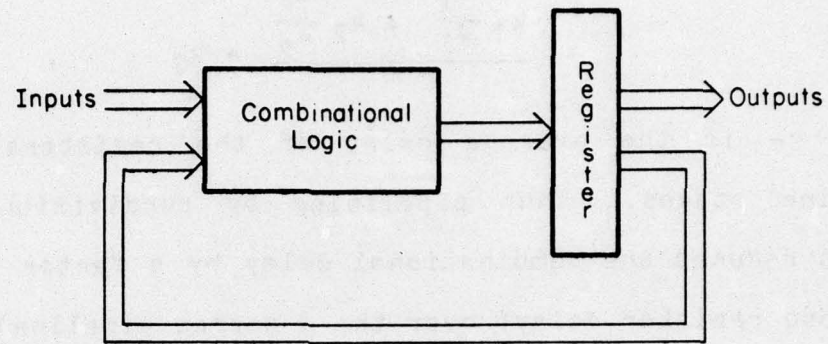
FP-5425

Figure 8. Finite State Machine Without Feedback
Before and After Pipelining

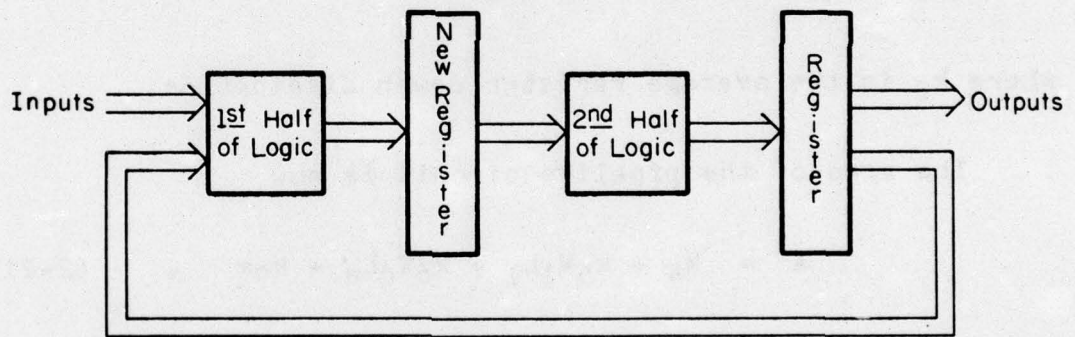
cases, the optimum clock interval is the maximum of the delays between the registers. In general, the system performance increases asymptotically toward a maximum value as the degree of pipelining increases. Simultaneously, the system cost increases in approximately linear fashion with an increase in pipeline degree.¹⁰

A finite-state machine with feedback is depicted in Fig. 9a. Here, pipelining by subdivision may be accomplished in the same manner as before. The combinational logic is cut at the half delay point and a register is inserted as in Fig. 9b. However, in this instance new operands may not necessarily be introduced at each clock interval. The circuit has a delay or time length of two clock intervals. Thus two separate tasks may be resident in the machine at any given time, with one task occupying each half of the circuit. New operands for each task are presented every other clock interval. As before, the finite state machine may be subdivided into m stages. In this case, new input operands for each task are presented once every m clock intervals.

The parametric equations for the m stage pipeline are slightly different from before. The speed of a pipelined circuit is defined as the effective throughput delay (e.g., the time between tasks rather than the delay of a single task). So, the time is set equal to the combinational delay of the non-pipelined case plus m times the register delay,



(a)



(b)

FP-5426

Figure 9. Finite State Machine with Feedback Before and After Pipelining

all divided by m .

$$T = \frac{k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d} + mk_7}{m} \quad (2-18)$$

$$= \frac{k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d}}{m} + k_7, \quad (2-19)$$

where k_7 is the average delay of the registers between pipeline stages. Thus pipelining by subdivision into m stages reduces the combinational delay by a factor of m and adds one register delay (over the 0 degree pipeline).

The power dissipation for a pipeline is

$$P = k_3 \frac{W_1}{L_1} + k_8 m, \quad (2-20)$$

where k_8 is the average register power dissipation.

The area of the pipeline circuit is now

$$A = k_4 + k_5 W_1 L_1 + k_6 W_d L_d + k_9 m, \quad (2-21)$$

where k_9 is the average register area.

The parametric equations for the different classes of circuits are summarized in Table 2.

Table 2.

Characterization of MOS Circuits

Function	Degree of Pipelining	
	<u>m=0</u>	<u>m≥1</u>
Speed	$k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d}$	$\frac{k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d}}{m} + k_7$
Power	$k_3 \frac{W_1}{L_1}$	$k_3 \frac{W_1}{L_1} + k_8 m$
Area	$k_4 + k_5 W_1 L_1 + k_6 W_d L_d$	$k_4 + k_5 W_1 L_1 + k_6 W_d L_d + k_9 m$

3. OPTIMUM CIRCUIT SPEED, POWER, AND AREA

Given the model and parametric equations developed in Chapter Two, it is now possible to investigate the dependence of MOS/LSI circuit speed, power, and area as functions of the channel dimensions and of the degree of pipelining. The conditions for optimizing each of the circuit functions individually is explored first. Then the effects of Beta ratio are examined.

3.1 Circuit Area

The area of an MOS/LSI IC is directly related to its cost. Fairchild Semiconductor has presented a qualitative estimate of the cost per unit area for MOS devices (Fig. 10).¹² This curve may be further approximated by a piecewise linear fit, also in Fig. 10. The left-hand segment has a slope near zero, so cost and area are related by a constant coefficient. Most MOS design occurs in this region, so this

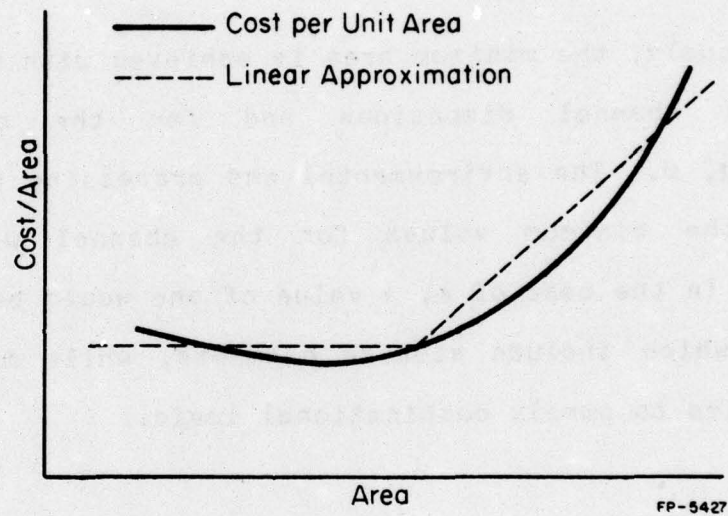


Figure 10. IC Cost per Unit Area as a Function of Area (from 12)

paper will simply be concerned with optimum area. For larger ICs, the right-hand portion of the curve suggests a more complex relationship between cost and area and is beyond the scope of this paper. But in either case, an objective of minimum cost implies minimum area. From Table 2,

$$A = k_4 + k_5 W_1 L_1 + k_6 W_d L_d + k_9 m \quad (3-1)$$

Obviously, the minimum area is achieved with minima for all four channel dimensions and for the degree of pipelining, m . The environmental and processing parameters dictate the minimum values for the channel widths and lengths. In the case of m , a value of one would be used for machines which include storage elements, while m equal to zero applies to purely combinational logic.

3.2 Circuit Power

The power dissipated on an IC, from Table 2, is

$$P = k_3 \frac{W_1}{L_1} + k_8 m \quad (3-2)$$

The minimum power dissipation results from a minimum for W_1 , a maximum for L_1 , and a minimum for m . $W_{1\min}$ and $L_{1\max}$ are set by the environmental and processing parameters, which are dictated to the designer. The value of m is either zero or one depending on the circuit in

question, as in the case of minimum area.

3.3 Circuit Speed

Also from Table 2, the speed of an IC is given by the formula

$$T = \frac{k_1 \frac{L_1}{W_1} + k_2 \frac{L_d}{W_d}}{m} + k_7 \quad (3-3)$$

The minimum time of the circuit is obtained with minima for L_1 and L_d , and maxima for W_1 , W_d , and m . As before, the bounds on the channel dimensions are pre-determined by processing and environmental parameters. For m , the maximum value is a function of the original logic. First, the circuit must be in a form which is pipelineable. The simplest pipelineable circuit is one that is entirely serial in nature. Tasks flow from one area of logic to the next with no feedback from a given area to a previous one. To pipeline such a serial circuit, registers may be inserted in the combinational areas between levels of logic. The maximum number of registers to be inserted is limited to the maximum number of logic levels between two successive registers in the original logic times the number of combinational sections. The time delay equation, 3-3, is valid only in the region $m=1$ to m equal to the upper bound just mentioned.

Many circuits which are not strictly serial are still pipelineable. Finite-state machines may be pipelined in much the same fashion as serial machines, as described in Chapter Two. Circuits which are combinations of several serial or finite-state machines are also sometimes pipelineable. For example, each section of a circuit which is a recognizable serial or finite-state machine may be pipelined independent of the rest of the circuit. Selective use of this technique could improve circuit throughput when applied to the "bottlenecks" or slowest sections. However, the designer must take care that the time dependencies between the sections of the circuit are maintained.

3.4 Beta Ratio Considerations

Each of the parametric equations - circuit speed, circuit area, and circuit power dissipation - have been optimized as functions of the four channel dimensions and the degree of pipelining. However, it is common industry practice to select both W_1 and L_d to take on their minimum values, $W_{1_{\min}}$ and $L_{d_{\min}}$, respectively. For the discussion of more complex optimality criteria in Chapter 4, W_1 and L_d will be considered as constants.

Several observations may be made concerning this choice. First, from Chapter 1, the steady-state requirements of a logic gate create a minimum acceptable Beta ratio. Recall that

$$\text{Beta} = \frac{W_d L_1}{W_1 L_d} \quad (3-4)$$

In general practice, the minimum required value of Beta is greater than that value achieved using the minima for the four channel dimensions which are dictated by the processing and environmental parameters:

$$\text{Beta}_{\min} \geq \frac{W_{d\min} L_{1\min}}{W_{1\min} L_{d\min}} \quad (3-5)$$

In order to achieve the necessary Beta, either W_d or L_1 or both must be increased above its minimum, even if the minimum values of W_1 and L_d are chosen. Of course, if W_1 or L_d is greater than its minimum, W_d or L_1 must be increased even further.

From Table 2, the equation for circuit power varies directly with W_1 and inversely with L_1 . Here, an increase in W_1 above its minimum would require a proportionate increase in either L_1 or W_d , so the power could at best remain the same, or at worst increase with the increase in W_1 .

From the area equation in Table 2, increases in either W_1 or L_d would increase total area. The accompanying increase in L_1 or W_d required by Beta_{\min} would further increase area.

So, for both power and area, it is obvious that the optimum choices for W_1 and L_d are their minima.

In the equation for speed in Table 2, both W_1 and W_d appear in denominators, with L_1 and L_d in numerators. An increase in L_d requires a proportionate increase in the product of L_1 and W_d . At best, this would leave speed unchanged, and at worst increase L_1 and L_d by the same amount, thus increasing the delay time.

Concerning W_1 , an increase above its minimum would require an increase in the product of L_1 and W_d . But here, if the increases were only applied to W_1 and to W_d , the minimum Beta bound would be satisfied and the delay time would decrease. In the extreme, this implies that it is possible to grow W_1 and W_d without bound in order to reduce the delay time. However, as a channel width increases beyond a certain limit, the simplified MOS model fails and the delay time actually begins to increase with increasing channel width.

In effect, the k_1 term in the speed equation is the only one which does not categorically support the choice of $W_{1\min}$ and $L_{d\min}$. For a particular case, with assigned values for the k_1 's, the simplification, fixing W_1 and L_d to their minimum values, should be reexamined.

Choosing $W_{1\min}$ and $L_{d\min}$ greatly simplifies the parametric equations. Let

$$k_1' = \frac{k_1}{W_{1\min}}, \quad (3-6)$$

$$k_2' = k_2 L_{d\min}, \quad (3-7)$$

$$k_3' = k_3 W_{1\min}, \quad (3-8)$$

$$k_5' = k_5 W_{1\min}, \quad \text{and} \quad (3-9)$$

$$k_6' = k_6 L_{d\min}. \quad (3-10)$$

With these substitutions, the equations appear as in Table 3.

Table 3.

Simplified Characterization of MOS Circuits

Function	Degree of Pipelining	
	<u>$m=0$</u>	<u>$m \geq 1$</u>
Speed	$k_1' L_1 + \frac{k_2'}{W_d}$	$\frac{k_1' L_1 + \frac{k_2'}{W_d}}{m} + k_7$
Power	$\frac{k_3'}{L_1}$	$\frac{k_3'}{L_1} + k_8 m$
Area	$k_4 + k_5' L_1 + k_6' W_d$	$k_4 + k_5' L_1 + k_6' W_d + k_9 m$

4. OPTIMA INVOLVING TWO CRITERIA

Although speed, power, and area are each of interest to the circuit designer, it is usually some combination of these that is the objective to be achieved. In this chapter, five objective functions are examined for the case of $m > 0$, i.e., sequential machines. These include the three pairwise products of the parametric equations - speed-power, power-area, and speed-area. Also, both minimum area and minimum power are derived assuming a required circuit speed. In addition, optima for the pairwise products are derived for $m = 0$, or purely combinational logic.

4.1 Speed And Power Optima

4.1.1 Speed-Power Product -

The speed-power product is a function mentioned often, especially when comparing various IC processes (TTL, ECL, MOS, etc.). As mentioned before, the speed equation actually represents a delay time. Although some references

are made in the literature to the power-delay product this paper uses the term speed-power product, which is more common. Also, pipelining, a factor not usually considered, is included in this analysis.

Multiplication of the power and speed equations gives

$$PT = \left(\frac{k_3'}{L_1} + k_8 m \right) \left(\frac{k_1' L_1 + \frac{k_2'}{W_d}}{m} + k_7 \right) \quad (4-1)$$

$$= k_1' k_8 L_1 + k_7 k_8 m + \frac{k_2' k_3'}{W_d L_1 m} + \frac{k_3' k_7}{L_1} + \frac{k_2' k_8}{W_d} + \frac{k_1' k_3'}{m} \quad (4-2)$$

From the form of equation 4-2 several observations can be made. First, the speed-power product is inversely related to W_d , so minimum PT implies maximum W_d .

Optimum values for m and L_1 may be obtained by taking partial derivatives:

$$\frac{\partial PT}{\partial m} = k_7 k_8 - \frac{k_2' k_3'}{W_d L_1 m^2} - \frac{k_1' k_3'}{m^2} \quad , \text{ and } \quad (4-3)$$

$$\frac{\partial PT}{\partial L_1} = k_1' k_8 - \frac{k_2' k_3'}{W_d m L_1^2} - \frac{k_3' k_7}{L_1^2} \quad . \quad (4-4)$$

These two partial derivatives, when set equal to zero, provide two simultaneous equations with two unknowns. However, substitution of one equation into the other results in a fifth order equation, for which no algebraic solution has yet been found.

It is still possible to derive optimum values for m and L_1 , assuming that one of them is fixed. Setting equation 4-4 equal to zero and solving for the positive value of L_1 yields

$$L_{1\text{opt}} = \sqrt{\frac{k_2'k_3' + k_3'k_7W_d m}{k_1'k_8W_d m}} \quad (4-5)$$

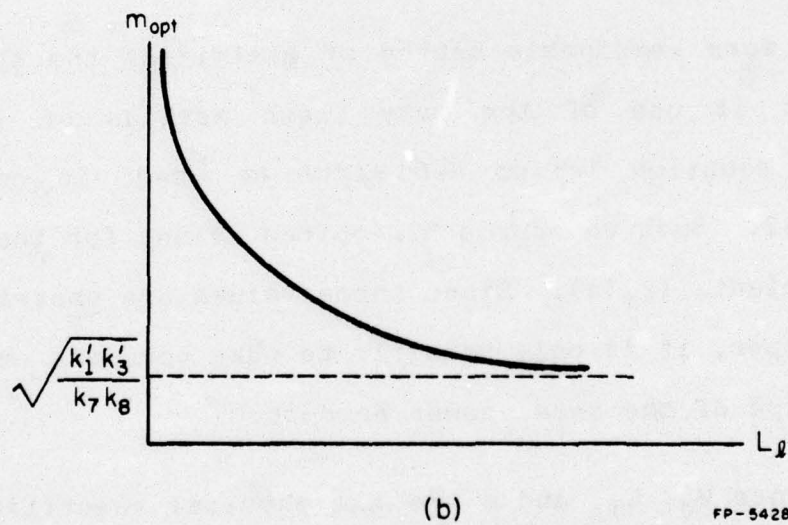
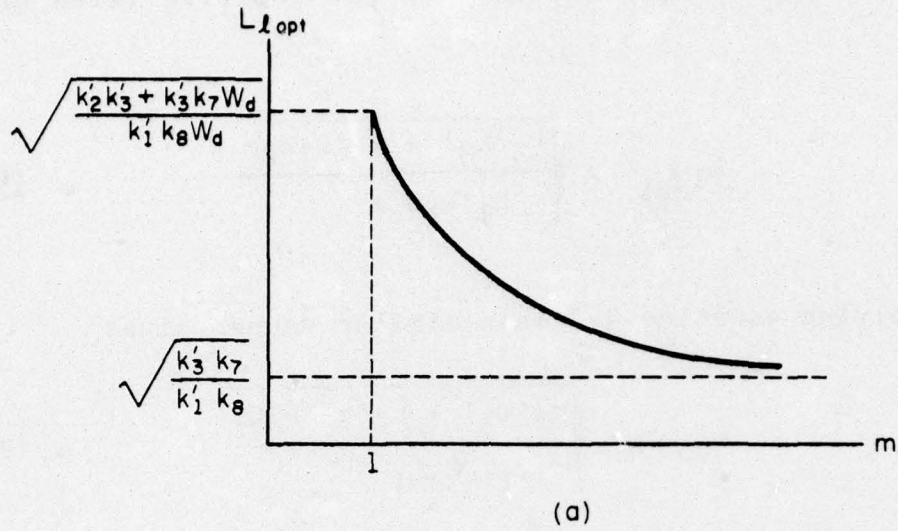
Solving equation 4-3 in a similar manner gives

$$m_{\text{opt}} = \sqrt{\frac{k_2'k_3' + k_1'k_3'W_d L_1}{k_7k_8W_d L_1}} \quad (4-6)$$

These two values are graphed in Fig. 11.

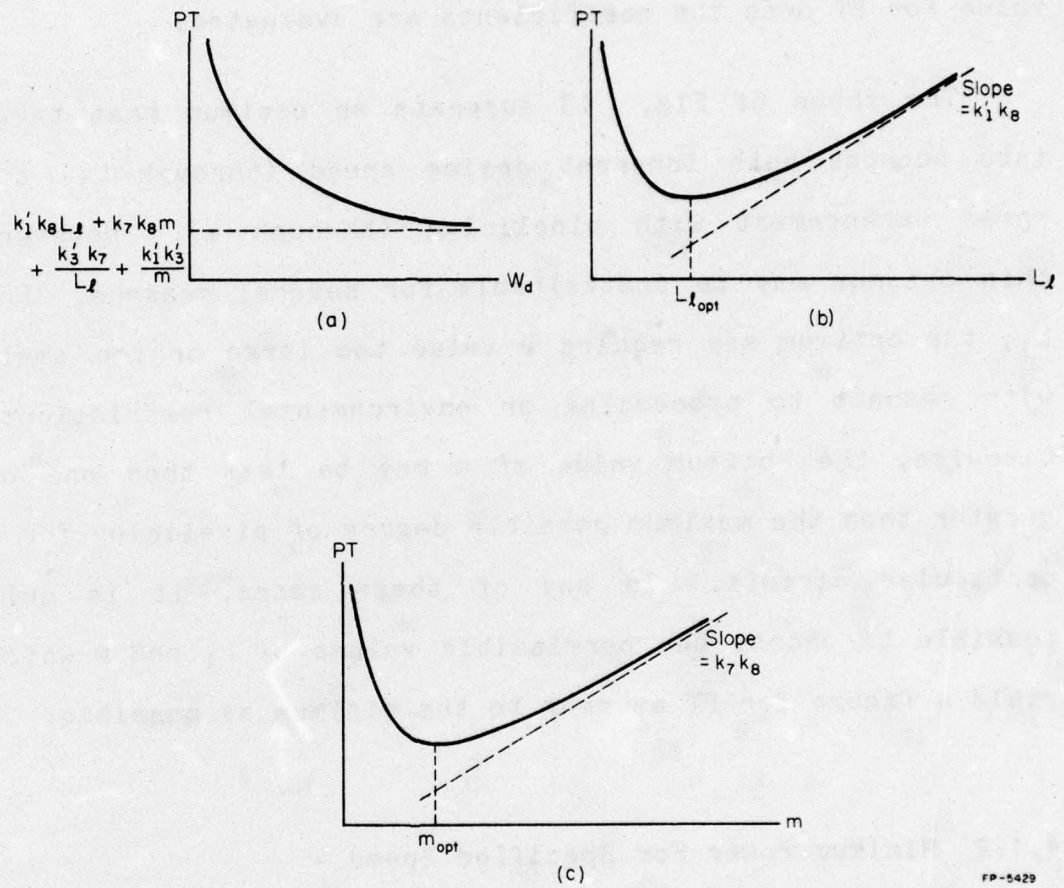
A more reasonable method of minimizing the speed-power product is one of the many known methods of iteration. Either equation 4-5 or 4-6 might be used, in conjunction with 4-2. Such an approach requires values for the constant coefficients (k_i 's). Since these values are unavailable for this paper, it is only possible to make comments on the form and shape of the speed-power product.

Since W_d , L_1 , and m are all physical quantities, it is only necessary to examine positive values for them. The general shape of PT with respect to W_d , L_1 , and m , individually, is depicted in Fig. 12. As mentioned before, the optimum value for W_d is a maximum, independent of the values of L_1 and m . The labels m_{opt} and $L_{1\text{opt}}$ are the values derived in equations 4-5 and 4-6.



FP-5428

Figure 11. L_{opt} and m_{opt} for Minimum Speed-Power Product



FP-5429

Figure 12. Speed-Power Product as a Function of W_d , L_f , and m , Individually

Optimum values for L_1 and m , however, are highly interdependent. In Fig. 13, the shape of PT is shown as a function of both L_1 and m . Since the shape is a hyperboloid, it is straightforward to iterate to the minimum value for PT once the coefficients are evaluated.

The shape of Fig. 13 suggests an optimum that takes into account both inherent device speed (through L_1) and speed enhancement with pipelining (through m). However, this optimum may be unattainable for several reasons. For L_1 , the optimum may require a value too large or too small with respect to processing or environmental restrictions. Likewise, the optimum value of m may be less than one or greater than the maximum possible degree of pipelining for a particular circuit. In any of these cases, it is only possible to choose the permissible values of L_1 and m which yield a figure for PT as near to the minimum as possible.

4.1.2 Minimum Power For Specified Speed -

Often, a circuit designer has a predetermined design goal for the speed of a circuit. This then becomes a constraint with respect to other design objectives. The particular objective function examined in this section is the optimization of power dissipation, assuming a bound on the circuit speed.

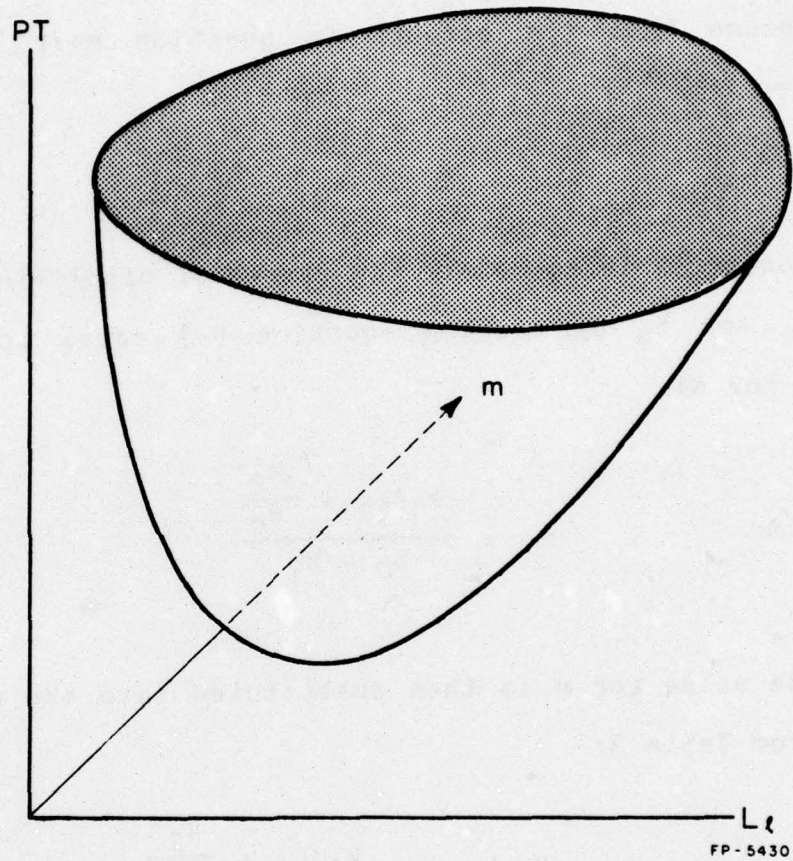


Figure 13. Speed-Power Product as a Function of both L_f and m

From Table 3, the speed of a circuit is given by the formula

$$T = \frac{k_1' L_1 + \frac{k_2'}{W_d}}{m} + k_7 \quad (4-7)$$

Now, assume that the circuit in question must meet the requirement

$$T \leq t_0 \quad (4-8)$$

It is now possible to relate the degree of pipelining, m , to L_1 , W_d , and t_0 by setting equation 4-7 equal to t_0 and solving for m :

$$m = \frac{k_1' L_1 + \frac{k_2'}{W_d}}{t_0 - k_7} \quad (4-9)$$

This value for m is then substituted into the equation for P from Table 3:

$$P = \frac{k_3'}{L_1} + k_8 \frac{k_1' L_1 + \frac{k_2'}{W_d}}{t_0 - k_7} \quad (4-10)$$

A maximum for W_d leads to a minimum for this equation. This equation may be further optimized by taking its partial derivative with respect to L_1 :

$$\frac{\partial P}{\partial L_1} = \frac{k_1' k_8}{t_0 - k_7} - \frac{k_3'}{L_1^2} \quad (4-11)$$

Setting this equation equal to zero and solving for L_1 ,

which must be positive, yields the result

$$L_{1\text{opt}} = \sqrt{k_3' \frac{t_0 - k_7}{k_1' k_8}} \quad (4-12)$$

This equation represents the optimum choice of L_1 in order to obtain minimum power dissipation, for a specified bound on speed, t_0 . This value for L_1 may then be substituted into equation 4-7, giving

$$m_{\text{opt}} = \frac{\sqrt{k_1' k_3' \frac{t_0 - k_7}{k_8}} + \frac{k_2'}{W_d}}{t_0 - k_7} \quad (4-13)$$

Equations 4-12 and 4-13 now provide the optimum values for L_1 and m as functions of a speed bound, t_0 , as shown in Fig. 14. The optimum L_1 increases monotonically proportional to the square root of $t_0 - k_7$. In other words, as the speed bound is relaxed (larger t_0 implies a slower requirement), L_1 is larger in order to achieve the minimum power. Likewise, m decreases monotonically with increasing t_0 (eventually proportional to the inverse of the square root of t_0), because the degree of pipelining required to achieve a slower speed is less.

In the case of minimum power with a speed bound, it is possible to calculate the optimum values for L_1 , W_d , and m algebraically. Iterative techniques are not necessary. These optimum values may then be substituted into the equation for power in Table 3 to provide the minimum power

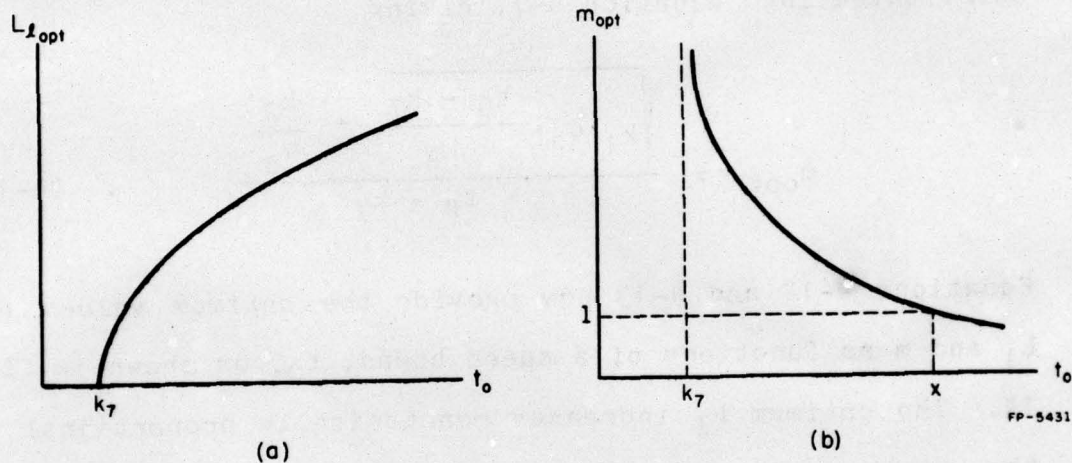


Figure 14. Optimum Values of L_ℓ and m as Functions of t_0 for Minimum Power

for the specified t_0 .

It is interesting to examine the case where the designer is restricted to a degree one pipeline, or $m=1$. This represents a simple finite state machine approach. From Fig. 14b, it is obvious that for t_0 greater than or equal to x , the optimum choice for m is indeed one. But for any choice of t_0 less than x , $m>1$ is optimum. So in this range, a restriction of $m=1$ will yield a value for power which is greater than the best obtainable.

4.2 Area-Power Product

Multiplication of the area and power equations yields the formula

$$AP = (k_4 + k_5 L_1 + k_6 W_d + k_9 m) \left(\frac{k_3'}{L_1} + k_8 m \right) \quad (4-14)$$

$$\begin{aligned} &= k_3' k_5' + \frac{k_3' k_4}{L_1} + \frac{k_3' k_6' W_d}{L_1} + \frac{k_3' k_9 m}{L_1} \\ &\quad + k_4 k_8 m + k_5' k_8 L_1 m + k_6' k_8 W_d m + k_8 k_9 m^2 \quad (4-15) \end{aligned}$$

A minimum for this equation requires minima for both W_d and m , as they appear only in numerators. To obtain the proper value for L_1 , consider

$$\frac{\partial AP}{\partial L_1} = k_5' k_8 m - \frac{k_3' k_4 + k_3' k_6' W_d + k_3' k_9 m}{L_1^2} \quad (4-16)$$

Setting equation 4-16 to zero and solving for L_1 , which must

be positive, yields the formula

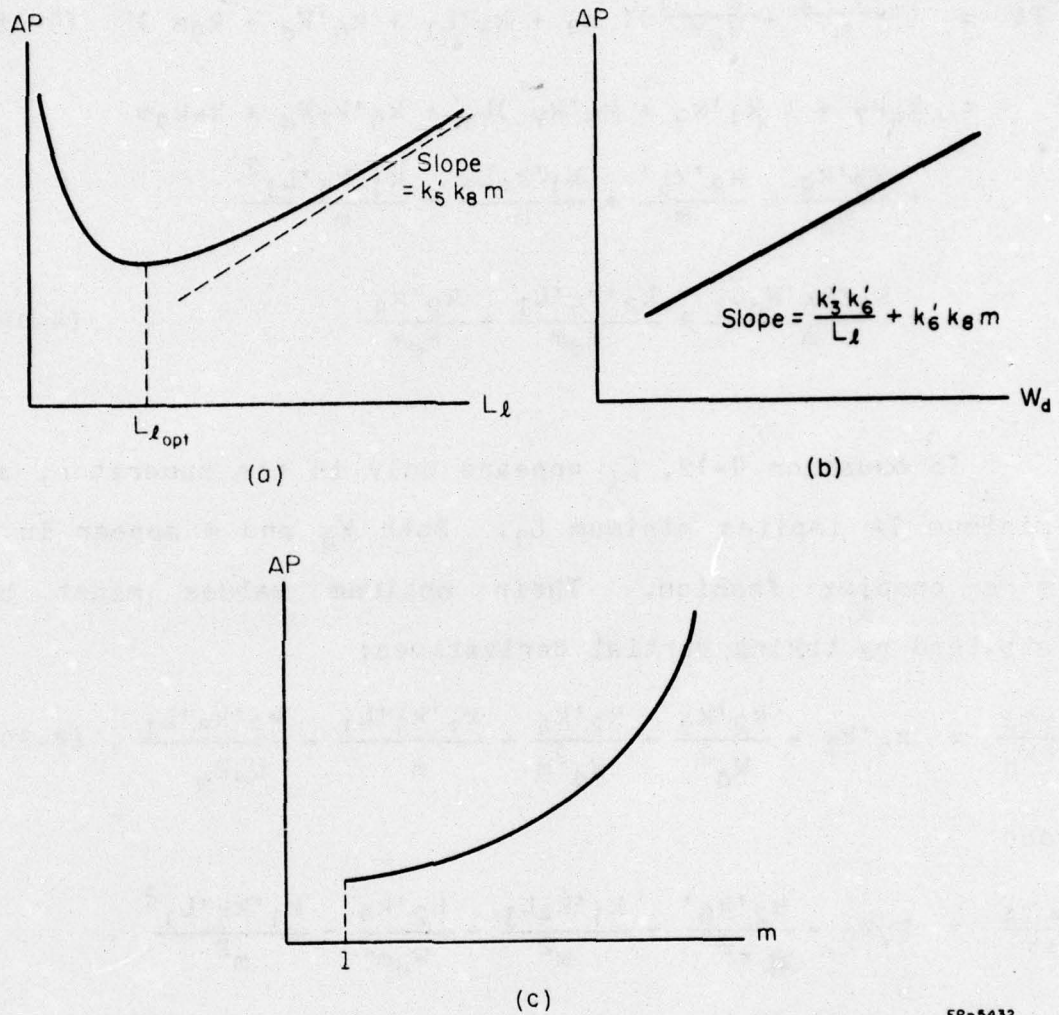
$$L_1 = \sqrt{\frac{k_3'k_4 + k_3'k_6'W_d + k_3'k_9m}{k_5'k_8m}} \quad (4-17)$$

The variation of the area-power product with respect to each of the three variables is demonstrated in Fig. 15. As mentioned before, the optimum values for L_1 , W_d , and m may not be physically exactly realizable. Interestingly, the restriction of $m=1$ yields the optimum value of AP, as shown in Fig. 15c. An increase in the degree of pipelining is useful to increase speed, which is not considered in this section.

4.3 Speed And Area Optima

4.3.1 Speed-Area Product -

The product of the speed and area equations is a design criterion of great interest. Often, the speed of a system is used to denote its performance. Also, in this case, area is related to cost. So, speed-area is actually a measure of a circuit's performance-cost ratio, since speed is characterized by its inverse, time. A minimum speed-area product is related to minimizing both time and cost, providing the maximum performance-cost ratio.



FP-5432

Figure 15. Variation of Area-Power Product with Respect to L_l , W_d , and m , Individually

Multiplication of the speed and area equations from Table 3 gives the equation

$$TA = \left(\frac{k_1' L_1}{m} + \frac{k_2' + k_7}{W_d m} \right) (k_4 + k_5' L_1 + k_6' W_d + k_9 m) \quad (4-18)$$

$$\begin{aligned} &= k_4 k_7 + (k_1' k_9 + k_5' k_7) L_1 + k_6' k_7 W_d + k_7 k_9 m \\ &\quad + \frac{k_2' k_9}{W_d} + \frac{k_2' k_6'}{m} + \frac{k_1' k_4 L_1}{m} + \frac{k_1' k_5' L_1^2}{m} \\ &\quad + \frac{k_1' k_6' W_d L_1}{m} + \frac{k_2' k_5' L_1}{W_d m} + \frac{k_2' k_4}{W_d m} \end{aligned} \quad (4-19)$$

In equation 4-19, L_1 appears only in the numerator, so minimum TA implies minimum L_1 . Both W_d and m appear in a more complex fashion. Their optimum values might be attained by taking partial derivatives:

$$\frac{\partial TA}{\partial W_d} = k_6' k_7 - \frac{k_2' k_9}{W_d^2} - \frac{k_2' k_4}{W_d^2 m} + \frac{k_1' k_6' L_1}{m} - \frac{k_2' k_5' L_1}{W_d^2 m}, \quad (4-20)$$

and

$$\begin{aligned} \frac{\partial TA}{\partial m} &= k_7 k_9 - \frac{k_2' k_6'}{m^2} - \frac{k_1' k_4 L_1}{m^2} - \frac{k_2' k_4}{W_d m^2} - \frac{k_1' k_5' L_1^2}{m^2} \\ &\quad - \frac{k_1' k_6' W_d L_1}{m^2} - \frac{k_2' k_5' L_1}{W_d m^2} \end{aligned} \quad (4-21)$$

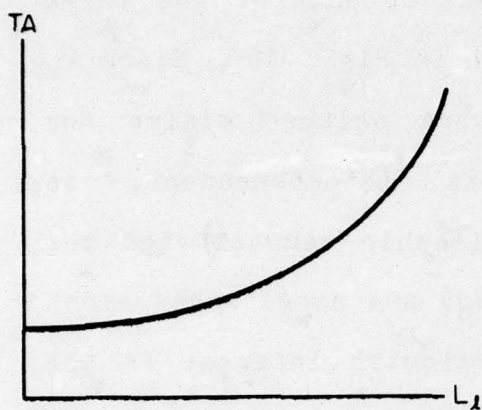
But, like optimum speed-power product, the optimum speed-area product does not lend itself to algebraic solution. However, optimum values for W_d and m may be obtained if one of them is considered fixed, following the example in section 4.1.1. When values for the various

constant coefficients are available, iteration to the minimum value of TA may be used.

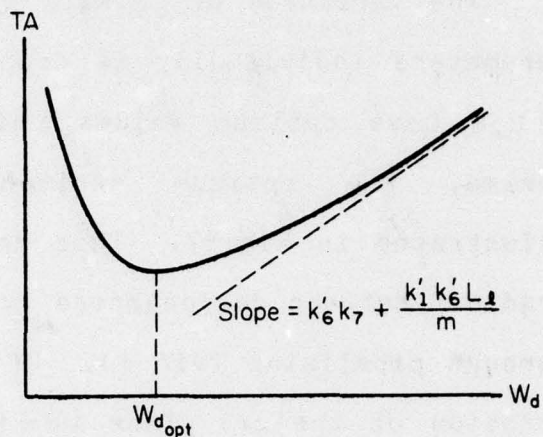
The variation of TA with respect to each of the three parameters individually is depicted in Fig. 16. Since W_d and m have optimum values which are neither minima nor maxima, the optimum choices are interdependent, as illustrated in Fig. 17. This relationship demonstrates the tradeoff between device speed (via W_d) and speed enhancement through pipelining (via m). Of particular interest is the location of the $m=1$ plane in Fig. 17, which specifies the available values of TA without pipelining. Several possibilities for the location of this plane exist.

If the $m=1$ plane intersects the hyperboloid at its minimum point, then the optimum performance-cost ratio may be obtained by simply choosing the appropriate value of W_d . If the minimum value for TA occurs at $m < 1$, then the optimum performance-cost ratio is unattainable, and $m=1$ should be chosen. Then the optimum value of W_d for $m=1$ should be selected.

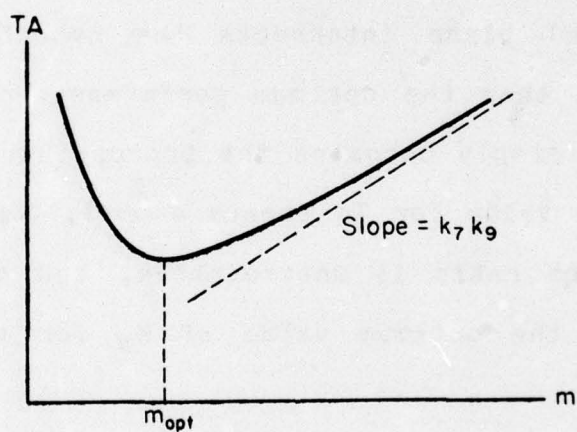
Finally, if the minimum value for TA occurs at $m > 1$, then an optimum performance-cost ratio may only be achieved with the aid of pipelining. A choice of $m=1$ restricts the value to the minimum point on the TA surface which intersects the $m=1$ plane.



(a)



(b)



(c)

FP-5433

Figure 16. Variation of TA with Respect to L_d , W_d , and m , Individually

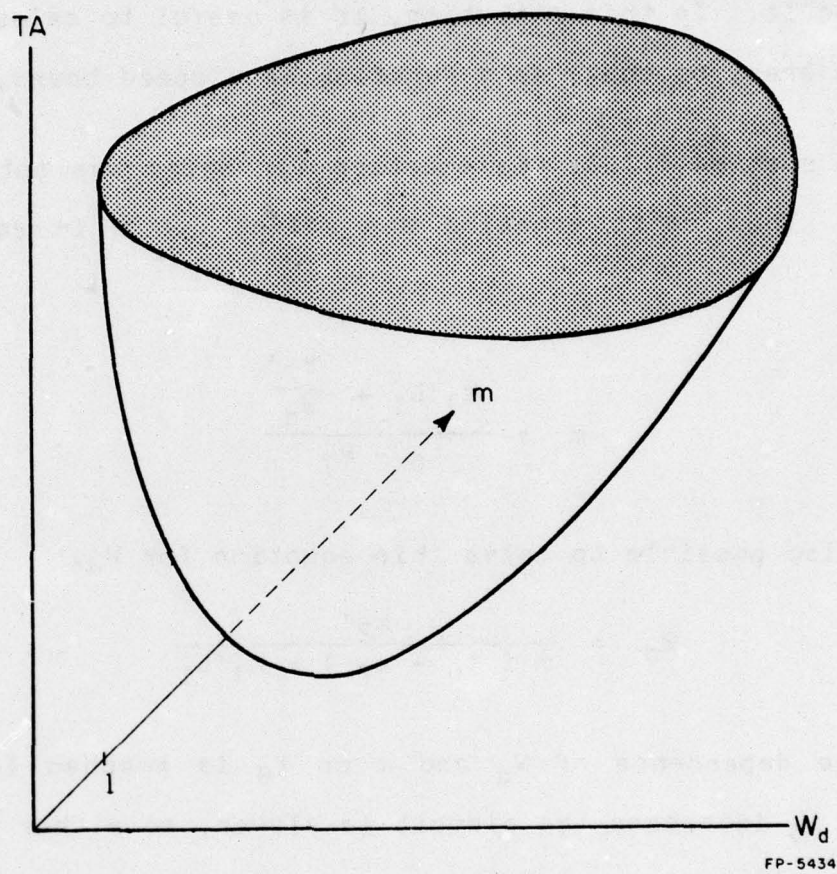


Figure 17. Speed-Area Product as a Function of Both m and W_d

4.3.2 Minimum Area For Specified Speed -

An optimum performance-cost ratio, as presented in the previous section, is not always desirable. There often exists a bound requirement on the performance, or speed, of the circuit. In this situation, it is useful to calculate a minimum area, or cost, as a function of a speed bound, t_0 .

In section 4.1.2, the equation for speed was set equal to t_0 . Then, this equation was solved for m in equation 4-7:

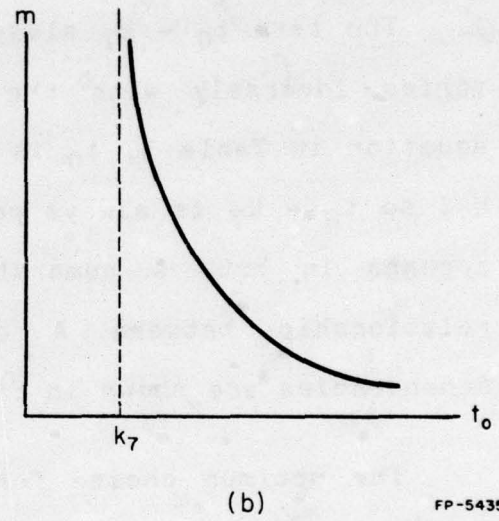
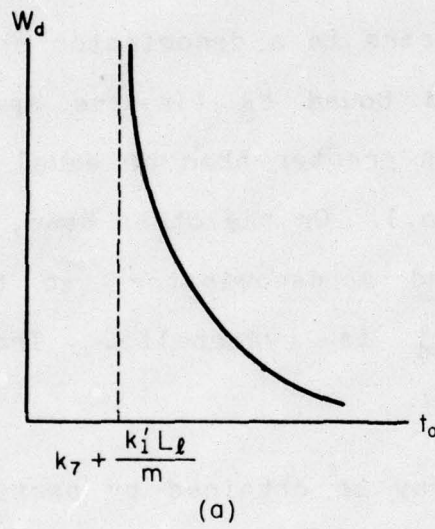
$$m = \frac{k_1' L_1 + \frac{k_2'}{W_d}}{t_0 - k_7} \quad (4-22)$$

It is also possible to solve this equation for W_d .

$$W_d = \frac{k_2'}{m (t_0 - k_7) - k_1' L_1} \quad (4-23)$$

Now, the dependence of W_d and m on t_0 is graphed in Fig. 18. As t_0 increases the circuit is slower, so either device speed or pipelining, or both, may be reduced.

To obtain a minimal area, either equation 4-22 or 4-23 may be substituted into the equation for area from Table 3. Using equation 4-22,



FP-5435

Figure 18. Variation of W_d and m as Functions of t_o

$$A = k_4 + k_5 L_1 + k_6 W_d + k_9 \frac{k_1' L_1 + \frac{k_2'}{W_d}}{t_0 - k_7} \quad (4-24)$$

$$= k_4 + k_5 L_1 + k_6 W_d + \frac{k_1' k_9 L_1}{t_0 - k_7} + \frac{k_2' k_9}{W_d (t_0 - k_7)} \quad (4-25)$$

As before, area is directly related to L_1 and minimum area implies a minimum for L_1 .

The term $t_0 - k_7$ always appears in a denominator so A varies inversely with the speed bound t_0 (In the speed equation in Table 3, t_0 is always greater than or equal to k_7 , so $t_0 - k_7$ is always positive.). On the other hand, W_d appears in both a numerator and a denominator, so the relationship between A and W_d is hyperbolic. These dependencies are shown in Fig. 19.

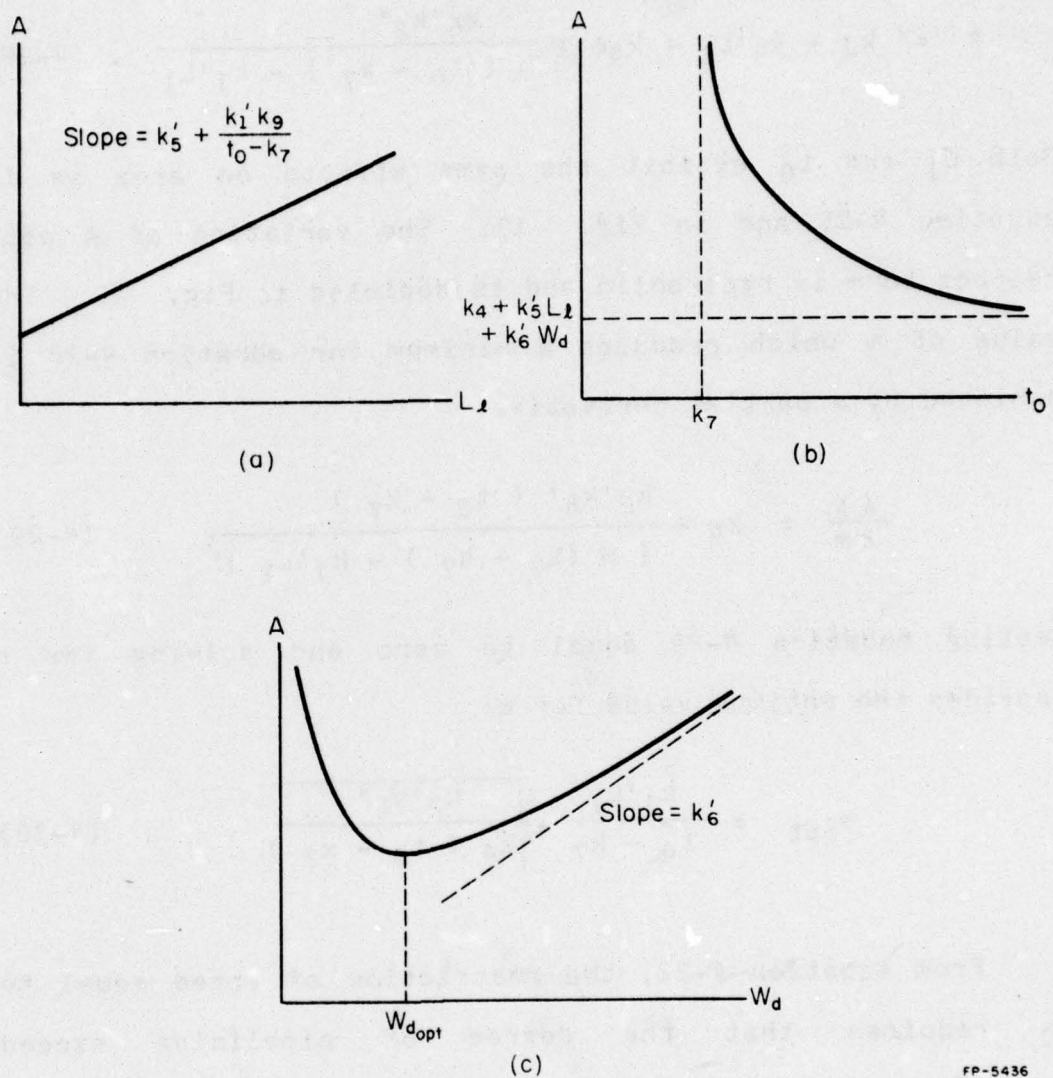
The optimum choice for W_d may be obtained by partial differentiation:

$$\frac{\partial A}{\partial W_d} = k_6' - \frac{k_2' k_9}{(t_0 - k_7) W_d^2} \quad (4-26)$$

Equation 4-26 is set equal to zero and solved for W_d , which must be positive, to yield

$$W_{d \text{ opt}} = \sqrt{\frac{k_2' k_9}{k_6' (t_0 - k_7)}} \quad (4-27)$$

Equation 4-27 is the value of W_d which minimizes A for a speed bound of t_0 .



FP-5436

Figure 19. Dependence of Area on L_l , t_o , and W_d Assuming

that $T = t_o$ and $m = \frac{k'_1 L_l + k'_2 / W_d}{t_o - k_7}$

Now, equation 4-23 is substituted into the area equation:

$$A = k_4 + k_5 L_1 + k_9 m + \frac{k_6' k_2'}{m (t_0 - k_7) - k_1' L_1} \quad (4-28)$$

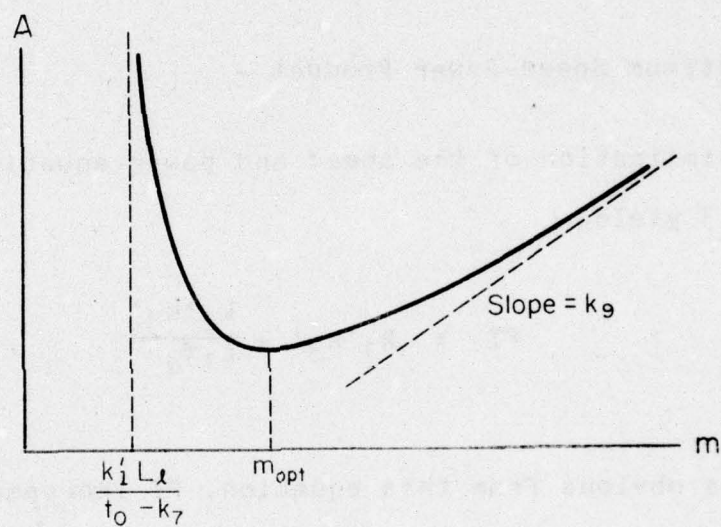
Both L_1 and t_0 exhibit the same effects on area as in equation 4-25 and in Fig. 19. The variation of A with respect to m is hyperbolic and is depicted in Fig. 20. The value of m which produces a minimum for equation 4-28 is achieved by a partial derivative:

$$\frac{\partial A}{\partial m} = k_9 - \frac{k_2' k_6' (t_0 - k_7)}{(m (t_0 - k_7) - k_1' L_1)^2} \quad (4-29)$$

Setting equation 4-29 equal to zero and solving for m provides the optimum value for m :

$$m_{opt} = \frac{k_1' L_1}{t_0 - k_7} + \sqrt{\frac{k_2' k_6'}{k_9 (t_0 - k_7)}} \quad (4-30)$$

From equation 4-22, the restriction of speed equal to t_0 requires that the degree of pipelining exceed $k_1' L_1 / (t_0 - k_7)$. Note that when m reaches this value, infinite area is required (see Fig. 20), corresponding to infinite W_d . If this value is greater than one, then the speed bound is unobtainable without a higher degree of pipelining. And, as before, even if $m=1$ is available as an option, it may severely increase area if m_{opt} in equation 4-30 is significantly greater than one.



FP-5437

Figure 20. Dependence of Area on m for $T = t_0$

4.4 Optima For Combinational Logic

The set of equations in Table 3 for $m=0$ describe the model for purely combinational logic. In this section, the pairwise products of these functions are optimized - speed-power, power-area, and speed-area under an $m=0$ constraint.

4.4.1 Optimum Speed-Power Product -

Multiplication of the speed and power equations for $m=0$ in Table 3 yields

$$PT = k_1'k_3' + \frac{k_2'k_3'}{L_1W_d} \quad (4-31)$$

As is obvious from this equation, PT increases linearly with reciprocals of L_1 and W_d . Therefore, PT is minimized by choosing maxima for L_1 and W_d . Of course, environmental and processing parameters place upper bounds on both of these channel dimensions.

4.4.2 Optimum Power-Area Product -

Equation 4-32 shows the power-area product for combinational logic.

$$AP = k_3'k_5' + \frac{k_3'k_4}{L_1} + \frac{k_3'k_6'W_d}{L_1} \quad (4-32)$$

AP grows linearly with W_d and with the reciprocal of L_1 . In this case the optimum L_1 is a maximum while the optimum W_d is a minimum.

4.4.3 Optimum Speed-Area Product -

The speed-area product is a measure of the performance cost ratio of the circuit. For strictly combinational logic, this product is:

$$TA = k_2'k_6' + k_1'k_4L_1 + k_1'k_5'L_1^2 + k_1'k_6'L_1W_d + \frac{k_2'k_4}{W_d} + \frac{k_2'k_5'L_1}{W_d} \quad (4-33)$$

TA increases linearly and quadratically with L_1 , so a minimum for L_1 is desired. With respect to W_d , however, neither a minimum nor a maximum is the optimum. As before, the optimum is achieved by taking a partial derivative and solving for the positive root:

$$W_{d\text{opt}} = \sqrt{\frac{k_2'k_4 + k_2'k_5'L_1}{k_1'k_6'L_1}} \quad (4-34)$$

This value for W_d minimizes TA for purely combinational logic.

Table 4 provides a summary of the results of the optimizations in chapters three and four.

Table 4.

Optimal Values of W_d , L_1 , and m

Function	L_1	W_d	m
<u>$m \geq 1$:</u>			
Area	min	min	min
Power	max	no effect	min
Speed	min	max	max
Speed-Power	intermediate*-A	max	intermediate*-A
Power-Area	$\sqrt{\frac{k_3'k_4+k_3'k_6'W_d+k_3'k_9m}{k_5'k_8m}}$	min	min
Speed-Area	min	intermediate*-B	intermediate*-B
Power($T=t_0$)	$\sqrt{\frac{k_3'(t_0-k_7)}{k_1'k_8}}$	max	$\sqrt{\frac{k_1'k_3'(t_0-k_7)}{k_8}} + \frac{k_2'}{W_d}$ (t_0-k_7)
Area($T=t_0$)	min	$\sqrt{\frac{k_2'k_9}{k_6'(t_0-k_7)}}$	$\frac{k_1'L_1}{t_0-k_7} + \sqrt{\frac{k_2'k_6'}{k_9(t_0-k_7)}}$
<u>$m=0$:</u>			
Area	min	min	
Power	max	no effect	
Speed	min	max	
Speed-Power	max	max	
Power-Area	max	min	
Speed-Area	min	$\sqrt{\frac{k_2'k_4+k_2'k_5'L_1}{k_1'k_6'L_1}}$	

* - No closed form, numerical solution easily obtained

A - Solve equations (4-3) = (4-4) = 0

B - Solve equations (4-20) = (4-21) = 0

5. NUMERICAL EXAMPLES

In order to illustrate the decisions and trade-offs in the design process further, two numerical examples are presented in this chapter. The equations used are those in Table 3 for $m > 0$, and the methods of optimization are detailed in Table 4.

The values chosen for the k_i 's are arbitrary, since the actual values depend both on the circuit under consideration and on proprietary process parameters. It should be noted that the particular values used were selected to illustrate certain points.

In addition, arbitrary bounds are selected for the parameters. L_1 will be allowed to range from 1 to 10 inclusive, and W_d from 0.1 to 5 inclusive. Also, the degree of pipelining, m , may vary between 1 and 5.

5.1 Example One

For this example, the k_i 's are assigned values as follows:

$$\begin{array}{lll} k_1' = 4 & k_4 = 10 & k_7 = 11 \\ k_2' = 1 & k_5' = 1 & k_8 = 1 \\ k_3' = 2 & k_6' = 2 & k_9 = 9 \end{array} \quad (5-1)$$

Using these values and Table 3 and Table 4, optimum values are calculated for speed, area, and time as well as their pairwise products. The results are presented in Table 5.

Each row of the table presents the equation which is optimized, the corresponding values of W_d , L_1 , and m , and the values of all the pertinent equations. The optimum value for each objective is circled for clarity.

In the cases of optimum speed-area(TA) and speed-power(PT) products, the desired value of m is less than one. Since a pipeline of degree one represents a finite-state machine, one is the minimum useable value for m in a finite-state machine. The lower section of Table 5 shows the optimum TA and PT for m restricted to one. This provides the minimum obtainable values in cases where the values computed are unrealizable.

The entry for optimum speed(T) with m restricted to one shows the best obtainable without pipelining, i.e., 15.2. This can be contrasted with the optimum with pipelining, 11.8. The maximum value for m is used because area and

Table 5.

Numerical Example 1

Optimized Function	m	W_d	L_1	A	P	T	TA	PT	AP
TA	0.815	0.84	1.0	20.0	2.82	17.4	347.	48.9	56.3
PT	0.862	5.0	2.37	30.1	1.71	22.2	670.	37.9	51.4
AP	1.0	0.1	6.2	25.4	1.32	45.8	1163.	60.6	33.6
A	1.0	0.1	1.0	20.2	3.0	25.0	505.	75.0	60.6
P	1.0	0.1	10.0	29.2	1.2	61.0	1781.	73.2	35.0
T	5.0	5.0	1.0	66.0	7.0	11.8	781.	82.9	462.
<u>m restricted to 1</u>									
TA	1.0	0.815	1.0	21.6	3.0	16.2	351.	48.7	64.9
PT	1.0	5.0	2.37	31.4	1.84	20.7	649.	38.1	57.8
T	1.0	5.0	1.0	30.0	3.0	15.2	456.	45.6	90.0

power are not considered. For this example, speed is the only equation whose optimum suffers by restriction to non-pipelining.

It is possible to solve for an objective which takes into account all three functions: speed, power, and area. For instance, one might wish to minimize the speed-area product under the constraint that power must not exceed some maximum value, P_{max} . The power equation from Table 3 is solved for one parameter, say L_1 . This formula is substituted for L_1 in equation 4-19. The resulting formula for TA is a function of variables W_d and m and the constant P_{max} . Iterative techniques are now applicable, as in Chapter 4. It should be noted that this procedure is only valid for those values of P_{max} which are less than that achieved when simply minimizing TA without regard to power, or 2.82 from Table 5. The results for minimum TA with a power bound are presented in Table 6. As the maximum power is decreased the minimum achievable TA increases.

5.2 Example Two

This example is presented to show an example in which pipelining is attractive. The k_i 's are assigned the following values:

Table 6.

Minimum TA for Specified Maximum P

P_{\max}	m	W_d	L_1	A	T	TA	PT	AP
2.75	1.0	0.804	1.14	21.8	16.8	366.	46.2	59.8
2.5	1.0	0.789	1.33	21.9	17.6	386.	44.0	54.8
2.25	1.0	0.769	1.6	22.1	18.7	414.	42.1	49.8
2.0	1.0	0.743	2.0	22.5	20.3	457.	40.7	45.0
1.75	1.0	0.707	2.67	23.1	23.1	533.	40.4	40.4
1.5	1.0	0.653	4.0	24.3	28.5	693.	42.8	36.5
1.25	1.0	0.56	8.0	28.1	44.8	1259.	56.0	35.2

$$\begin{array}{lll}
 k_1' = 6 & k_4 = 10 & k_7 = 4 \\
 k_2' = 2 & k_5' = 2 & k_8 = 1 \\
 k_3' = 2 & k_6' = 4 & k_9 = 3 \quad (5-2)
 \end{array}$$

Using the equations of Table 3 and the methods of Table 4, the results are generated and are shown in Table 7. In this example all functions dealing with speed (TA, PT, and T) benefit from pipelining.

In the case of optimum TA, pipelining provides a 27 percent decrease over non-pipelining. The decrease for PT is almost 6 percent, while T decreases by 49 percent. The trade-off is essentially between the delay in combinational logic and the area and power required by additional registers. For speed, T, the maximum allowable figure for m is chosen because area and power are not considered. Incidentally, the optimum values for TA and PT in Table 6 are the best obtainable with integer values of m. The equations yield optimum m values of 3.29 and 1.78, respectively.

As with example one, it is possible to find the minimum speed-area product(TA) subject to an upper bound on power, P_{\max} . In this case the value of P_{\max} must stay less than or equal to 5.0, from Table 7. The minimum TA is shown in Table 8 for various values of P_{\max} . As before, the optimum TA increases as the power bound is tightened or lowered. However, minimum TA does not change monotonically because L_1 is bounded by 1 and the optima using P_{\max} values of 4.5 and

Table 7.

Numerical Example 2

Optimized Function	m	W_d	L_1	A	P	T	TA	PT	AP
TA	3.0	0.765	1.0	24.1	5.0	6.87	165.	34.3	120.
PT	2.0	5.0	1.18	38.4	3.69	7.74	297.	28.6	142.
AP	1.0	0.1	3.67	20.7	1.54	46.0	954.	71.1	32.0
A	1.0	0.1	1.0	15.4	3.0	30.0	462.	90.0	46.2
P	1.0	0.1	10.0	33.4	1.2	84.0	2806.	101.	40.1
T	5.0	5.0	1.0	47.0	7.0	5.28	248.	37.0	329.
<u>m restricted to 1</u>									
TA	1.0	0.865	1.0	18.5	3.0	12.3	227.	36.9	55.4
PT	1.0	5.0	1.21	35.4	2.65	11.7	413.	30.3	94.0
T	1.0	5.0	1.0	35.0	3.0	10.4	364.	31.2	105.

Table 8.

Minimum TA for Specified Maximum P

P_{\max}	m	W_d	L_1	A	T	TA	PT	AP
5.0	3.0	0.764	1.0	24.1	6.87	165.	34.4	120.
4.5	3.0	0.736	1.33	24.6	7.57	186.	34.1	111.
4.0	2.0	0.802	1.0	21.2	8.25	175.	33.0	84.8
3.5	2.0	0.764	1.33	21.7	9.31	202.	32.6	76.0
3.0	1.0	0.866	1.0	18.5	12.3	227.	36.9	55.4
2.5	1.0	0.808	1.33	18.9	14.5	274.	36.2	47.2
2.0	1.0	0.729	2.0	19.9	18.7	373.	37.5	39.8
1.5	1.0	0.612	4.0	23.4	31.3	733.	46.9	35.2

3.5 occur with L_1 less than 1. In effect, increasing L_1 above the optimum choice forces m , an integer, to change value which in turn forces TA to take on a higher minimum value. In addition, the degree of pipelining, m , is forced lower by the decreasing power bound.

6. CONCLUSION

In this paper, a simplified model was devised for the logic in MOS/LSI circuits. First, a single MOS inverter was modeled, and its speed, power dissipation, and area were expressed as functions of transistor channel dimensions. This model was then scaled to represent the combinational logic of an entire circuit. The channel dimensions in the equations became the effective or average dimensions for the circuit as a whole. A register model was introduced and the circuit model was extended to reflect the effects of pipelining. Then, the speed, power, and area of the circuit were formulated using the parameters of the model: channel widths and lengths and the degree of pipelining. These equations were expressed with unevaluated constant coefficients which depend on proprietary industry information and on the specific circuit being analyzed.

The equations describing the circuit were then individually optimized with respect to the parameters. Minimum area results from minimization of all of the parameters. Minimum power requires maximum load channel length, minimum load width, and minimum pipelining and was independent of the driver channel dimensions. Optimum speed (minimum time) requires minimum channel lengths, maximum channel widths, and maximum pipelining.

After these observations, the model was further simplified to facilitate the study of more complex and more interesting optima. In this model, the circuit functions were dependent only on load channel length, driver channel width, and degree of pipelining, with load channel width and driver channel length set to their minimum values.

The first function considered was the product of speed and power. Although a closed form solution was not obtainable for the optimal parameter values, each optimum parameter value was expressed in terms of the other parameters. These equations could then be used to iterate to the minimum for the speed-power product. This procedure will succeed since the shape of PT is a hyperboloid. In addition, power was minimized assuming a bound on speed. In this case, the optimum values for the parameters were derived.

Next, the product of area and power was considered. Here, optimal values for all three parameters were attained.

Optima involving speed and area were dealt with next. The minimum speed-area product, like speed-power, could not be solved in a closed form. However, since it was also a hyperboloid, iterative techniques would succeed. Also, area was minimized for a bound on speed, and optimal values were obtained for all parameters.

For those circuits which are purely combinational, optimal parameters values for single and pairwise products of the objective functions were also derived and tabulated.

Finally, two numerical examples were presented to illustrate the methods involved. The unevaluated coefficients were assigned arbitrary values.

The models and circuit functions derived are useful in describing the nature of the various trade-offs and sacrifices inherent in MOS/LSI circuit design. With them, it can be shown whether a circuit would benefit from particular design choices, involving various channel dimensions and the use of pipelining. If the constant coefficients are evaluated using the knowledge of a MOS/LSI semiconductor manufacturer, then near optimum design choices may be made.

REFERENCES

1. Penney, William M., and Lillian Lau, ed. MOS Integrated Circuits. New York: Van Nostrand Reinhold Company, 1972.
2. Carr, William N., and Jack P. Mize. MOS/LSI Design and Application. New York: McGraw-Hill Book Co., 1972.
3. Grove, A.S. Physics and Technology of Semiconductor Devices. New York: John Wiley and Sons, Inc., 1967.
4. Penney and Lau, op. cit., p. 75.
5. Ibid., pp. 186-251.
6. Ibid., pp. 252-330.
7. Carr and Mize, op. cit..
8. Ibid..
9. Penney and Lau, op. cit..
10. Larson, A.G., and E.S. Davidson. "Cost Effective Design of Special Purpose Processors: A Fast Fourier Transform Case Study." Proc. Eleventh Annual Allerton Conference on Circuit and System Theory, October, 1973, pp. 547-557.
11. Shar, L.E., and E.S. Davidson. "A Multiminiprocessor System Implemented Through Pipelining." Computer, February, 1974, pp. 42-51.
12. Fairchild Semiconductor. OPTIMOS. Mountain View: Fairchild Camera and Instrument Corporation, 1972, pp. 6-7.